DOCUMENT RESUME

ED 080 545 TM 003 050

TITLE	Proceedings of the 1970 Invitational Conference on Testing Problems (New York, October 31, 1970).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	31 Oct 70
NOTE	184p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Conference Reports; Educational Change; *Information Dissemination; Information Needs; *Information Systems; Program Evaluation; Social Change; Statistical Analysis; *Testing Problems; *Test Results

ABSTRACT

The theme of the 1970 Invitational Conference on Testing Problems was "The Promise and Perils of Educational Information Systems." The following papers were presented: (1) "Longer Education: Thinner, Broader, or Higher" by Fritz Machlup; (2) "Testing: Americans' Comfortable Panacea" by Theodore R. Sizer; (3) "Social and Cultural Change and the Need for Educational Information: The Futurist's View" by Herman Kahn; (4) discussion of the previous papers by Amitai Etzioni; (4) "School Testing to Test the Schools" by Richard M. Jaeger; (5) "National Assessment" by Robert E. Stake; (6) discussion by James N. Jacobs and Frank B. Womer; (7) "Bayesian Considerations in Educational Information Systems" by Melvin R. Novick; (8) discussion by John W. Tukey; (9) "Temporal Changes in Treatment-Effect Correlations; A Quasi-Experimental Model for Institutional Records and Longitudinal Studies" by Donald T. Campbell; (10) discussion by John W. Tukey; (11) "Higher Education: For Whom? At Whose Cost?" by Carl Kaysen; (12) "Social Accounting in Education; Reflections on Supply and Demand" by David K. Cohen; (13) "Ethical and Legal Aspects of the Collection and Use of Educational Information" by David A. Goslin; (14) "Test Information as a Reinforcer of Negative Attitudes Toward Black Americans" by Elias Blake Jr.; and (15) discussion by James J. Gallagher. (KM)

FILMED FROM BEST AVAILABLE COPY

ED 080545

TM 003 050

U S DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Proceedings
of the

1970
Invitational Conference
on
Testing Problems

◆

The Promise and Perils
of
Educational Information
Systems

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Dorothy Urban
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

EDUCATIONAL TESTING SERVICE

FILMED FROM BEST AVAILABLE COPY

Copyright © 1971 by Educational Testing Service. All rights reserved.
Library of Congress Catalog Number: 47-11220
Printed in the United States of America

ED 080545

Proceedings
of the

1970
Invitational Conference
on
Testing Problems

October 31, 1970
The New York Hilton
New York City

Gene V Glass
Chairman

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY • BERKELEY, CALIFORNIA • EVANSTON, ILLINOIS

ETS
Board of Trustees
1970–71

Caryl P. Haskins, *Chairman*

Melvin W. Barnes
George F. Baughman
John Brademas
John T. Caldwell
Arland F. Christ-Janer
Charles C. Cole Jr.
John H. Fischer
Hugh M. Gloster
Samuel B. Gould
William J. McGill
Malcolm C. Moos
Paul C. Reinert, S.J.
William W. Turnbull
Albert N. Whiting
Logan Wilson

ETS Officers

William W. Turnbull, *President*
Robert J. Solomon, *Executive Vice President*
David J. Brodsky, *Vice President and Treasurer*
Henry S. Dyer, *Vice President*
John S. Helmick, *Vice President*
Richard S. Levine, *Vice President*
Winton H. Manning, *Vice President*
Samuel J. Messick, *Vice President*
Charles E. Scholl, *Vice President*
G. Dykeman Sterling, *Vice President*
Joseph E. Terral, *Vice President*
Catherine G. Sharp, *Secretary*
Russell W. Martin Jr., *Assistant Treasurer*
Herman F. Smith, *Assistant Treasurer*

Foreword

During times of social stability men gathered in conferences tend to accept the workings of the system and to report on their individual achievements. In times of social change, however, the mood changes. Then the criticism quotient rises—and for good reason. In such times, the system itself must be examined critically, both to discern which of the old ways should be discarded and which retained, and which of the new ways are to be encouraged and advanced.

Perhaps more than at any other time in the history of the Invitational Conference on Testing Problems it could be said that the prevailing mood of this year's Conference was one of concern about the systems of education and of measurement. As these papers make apparent, there is much to be done yet to bring our lagging institutions in line with a social reality that everywhere is pressing ahead. There is, I believe, something very hopeful in the ability of both the speakers and the audience at this Conference to face hard issues squarely and to explore candidly ideas that for some may be unconventional and even uncomfortable. The participants would perhaps agree with Oscar Wilde's dictum that "an idea that isn't dangerous is hardly worth calling an idea at all."

Conference Chairman Gene V Glass briefly summarizes the contribution of all the speakers in his preface. I would only note here the good fortune of those attending a program that opened with a critique by the internationally known economist Fritz Machlup, who challenged both the old and the new conventional wisdom about higher education, and ended with Elias Blake's eloquent assertion of the right of all Americans to more than token access to all forms of education and to the benefits thereof.

Much of the credit for this program must go to Dr. Glass as chairman. Social research indicates that solving complex problems often requires the attack of diverse as well as brilliant minds; certainly Dr. Glass's choice of speakers reflects such a favoring principle at work. We are grateful to him and to the speakers and discussants of the 1970 Invitational Conference for a contribution that may help to make a difference in the future.

William W. Turnbull
PRESIDENT

Preface

The theme of the 1970 Invitational Conference on Testing Problems was "The Promise and Perils of Educational Information Systems." Educational information systems are significant new phenomena in the world of schooling. These systems are collections of test data on knowledges, skills, interests, and attitudes of children and adults and are maintained for the purpose of educational decision making. The National Assessment of Educational Progress (NAEP) is very likely a prototype of educational information systems that will be developed in the 1970s by state and federal governmental agencies, educational institutions, and perhaps even international agencies. Currently several state departments of education (among them Michigan, Pennsylvania, New York, and Colorado) are developing assessment systems patterned more or less after NAEP. Whatever the future holds for educational information systems, it will undoubtedly be characterized by a plurality of such systems with diverse purposes and uses. The development of these systems poses questions that social scientists, educators, statisticians, and philosophers must address.

The creation of an educational information system raises both hopes and fears. The promise of more informed decision making, which resides in these newly created systems, is quickly tempered in the minds of thoughtful men by the realization that these powerful inventions can be harmful if used carelessly. Daniel P. Moynihan in *Maximum Feasible Misunderstanding* recalled John Kenneth Galbraith's observation of "the indispensable role of statisticians in modern societies, which seem never to do anything about problems until they learn to measure them, that being the special province of those applied mathematicians. Statistics are used as mountains are climbed: because they are there." Testing systems can signal the existence of problems currently unrecognized. From heightened self-consciousness can come better schooling. But statistics once gathered will find new uses, uses unanticipated by those who designed the system, and nothing guarantees wise utilization.

Writing in "The Learning Process in the Dynamics of Total Societies" (in *The Study of Total Society*), Kenneth Boulding noted that "We have been fairly successful in collecting and processing economic data on the scale of the total society, as the development of national income statistics proves. If we can structure the process on a regular, systematic, month-by-month basis for other essential social variables, it will constitute an enormous step forward towards a viable social science." However, the promise is not without a countervailing peril, as Boulding went on to point out: "All decisions are made on the basis of some image of the world derived from information processing. If, therefore, we introduce the collection and processing of social scientific information into the social system, we cannot expect it to remain unchanged, and the political sensitivity of such information collection and processing depends on this fact." All that we know for certain is that educational information systems possess the potential to change education. Whether the changes will be good or bad can only be seen now by men of wisdom and foresight.

The group of scholars who turned their thoughts to the promise and perils of educational information systems during the 1970 Invitational Conference can best be described in a word as "diverse": a futurist, an historian, an economist, a political scientist, a sociologist, a neo-reformation activist, an unashamed classicist. No allegation of professional narcissism in the program of the Invitational Conference can be made to stick. At the end of the 1946 Invitational Conference, a participant remarked that "for a long time this group has regarded itself as test technicians. The group is beginning to show a little more interest in the whole science of education." He was not so much marking a sharp break with the past as he was perceiving a slow trend that may be entering its final stages in the program of the Conference now nearly a quarter of a century later. The program of the 1970 Invitational Conference reflects a realization of the pervasive social consequences of the phenomenal inventions of twentieth-century psychometricians.

The first general session of the Conference was addressed to the educational and social context of educational information systems. Fritz Machlup, Walker Professor of Economics and International Finance at Princeton University, presented with refreshing forthrightness and candor a provocative opinion on the growth of higher education in the next decade. Theodore R. Sizer, Dean of the Harvard Graduate School of Education, examined the interplay of testing and social change throughout the history of the first half of twentieth-century America. Herman Kahn, Director of the Hudson Institute, projected the broad social trends that will shape, and to a lesser extent be shaped by, the educational information systems of the next three decades.

In the first of two concurrent sessions making up the second segment of the Conference, Robert E. Stake, Associate Director of the Center for Instructional Research and Curriculum Evaluation, University of Illinois, presented a critical analysis of the National Assessment of Educational Progress. Richard M. Jaeger, Chief of Evaluation Methodology, Bureau of Elementary and Secondary Education, USOE, continued the emphasis in this session on the realities of extant educational information systems with his examination of uses of large-city school testing programs. Concurrently, Melvin R. Novick of the American College Testing Program and University of Iowa described the application of recent advances in Bayesian statistics to a particular type of information system. The emphasis on technical issues in the latter concurrent session was extended by Donald T. Campbell of Northwestern University, who imaginatively explored some of the experimental purposes to which educational information systems could be applied.

During the luncheon Carl Kaysen, Director of the Institute for Advanced Study, projected the hard choices facing higher education in terms of population and economics.

In the second general session, attention was focused on the social and political problems that arise with the possibility of large-scale educational information systems. David K. Cohen, Executive Director of the Center for Educational Policy Research, Harvard Graduate School of Education, saw with acuity the political issues that will play an important role in the legitimation and utilization of testing systems. David A. Goslin, Russell Sage Foundation, refused to allow the testing fraternity to avoid addressing the moral, ethical, and legal questions posed by operational information systems. With a concern for pressing, contemporary social issues, Elias Blake Jr., President of the Institute for Services to Education, confronted the profession with the undeniable observation that educational testing is too often an indirect expression of one set of human values at work in a system dedicated to the protection of a plurality of values.

The Invitational Conference exists to bring before the profession the thinking of scholars such as these. All credit is due to them. A debt of gratitude is owed to those who kindly consented to act as discussants at the Conference and to prepare their reactions for this publication: Amitai Etzioni of Columbia University; James N. Jacobs, Cincinnati Public Schools; Frank B. Womer, NAEP and The University of Michigan; John W. Tukey, Princeton University; and James J.

Gallagher, University of North Carolina.

The Chairmanship of the Invitational Conference on Testing Problems is partly honorific, partly functional. The honor is humbling and, in the words of a former Chairman, causes one to remember "that others more deserving have not yet been so generously recognized." The burden of this Chairman's duties was lightened by several persons. John L. Hayman of the Great Cities Research Council assisted by chairing one of the concurrent sessions during the second portion of the Conference. During the development of the theme and slate of speakers for the Invitational Conference, the advice of the following ETS personnel was most courteously and generously offered: Henry Dyer, Richard Levine, Samuel Messick, Robert Solomon, and William Turnbull. Other ETS staff assisted in extending invitations to speakers. Finally, particular thanks are due to two members of the ETS staff: Kay Sharp, whose special talent for organizing the hundreds of detailed arrangements for these conferences created the well-ordered yet gracious atmosphere of the meeting, and Anna Dragositz who, from beginning to end, year after year, holds the Invitational Conference together with unerring professional judgment.

Gene V Glass
CHAIRMAN

EDUCATIONAL TESTING SERVICE

Measurement Award
1970

* * *

The ETS Award for Distinguished Service to Measurement was established in 1970, to be presented annually to an individual whose work and career has had a major impact on developments in educational and psychological measurement. The first of this new series of awards was presented during the Conference by ETS President William W. Turnbull to Professor E. F. Lindquist with the following citation:

> A man of rare foresight, ingenuity, and energy, possessing both the profound understanding of educational measurement and creative ideas for its application, E. F. Lindquist in a distinguished career for more than 40 years has given educational measurement new insights, new instruments, new techniques, and new technologies.
>
> As a result of his early and continuing interest in improving the measurement of academic potential and achievement, Dr. Lindquist developed tests which have become model ones, used today by schools and colleges throughout the United States and in English-speaking countries around the world. His concern for the effective collection and analysis of information led to Dr. Lindquist's invention of the first high-speed electronic scoring machines and supplementary devices that have vastly reduced the time required to process test papers and interpret the results.
>
> Through his teaching and research, his numerous articles, his several widely used texts, and his professional associations, Dr. Lindquist has enriched our understanding of statistical methods, of measurement theory, and especially of education itself.
>
> For his contributions to the scholarship and uses of educational measurement, Educational Testing Service is pleased to award the first ETS Award for Distinguished Service to Measurement to E. F. Lindquist.

E. F. Lindquist

Contents

iii Foreword by William W. Turnbull

iv Preface by Gene V Glass

viii Presentation of 1970 Measurement Award to E. F. Lindquist

General Session: Educational and Social Context

3 Longer Education: Thinner, Broader, or Higher,
Fritz Machlup, Princeton University

14 Testing: Americans' Comfortable Panacea,
Theodore R. Sizer, Harvard University

22 Social and Cultural Change and the Need for Educational
Information: The Futurist's View, Herman Kahn,
Hudson Institute

34 Discussion: Amitai Etzioni, Columbia University and
The Center for Policy Research

Session A: Educational Applications

39 School Testing to Test the Schools,
Richard M. Jaeger, U.S. Office of Education

53 National Assessment, Robert E. Stake, University of Illinois

67 Discussion: James N. Jacobs, Cincinnati (Ohio) Public Schools

71 Discussion: Frank B. Womer,
National Assessment of Educational Progress and
The University of Michigan

Session B: Technical Issues

77 Bayesian Considerations in Educational Information Systems, Melvin R. Novick, American College Testing Program and University of Iowa

89 Discussion: John W. Tukey, Princeton University

93 Temporal Changes in Treatment-Effect Correlations: A Quasi-Experimental Model for Institutional Records and Longitudinal Studies, Donald T. Campbell, Northwestern University

111 Discussion: John W. Tukey, Princeton University

Luncheon Address

114 Higher Education: For Whom? At Whose Cost? Carl Kaysen, Institute for Advanced Study

General Session: Problems in Implementation

129 Social Accounting in Education: Reflections on Supply and Demand, David K. Cohen, Harvard University

149 Ethical and Legal Aspects of the Collection and Use of Educational Information, David A. Goslin, Russell Sage Foundation

160 Test Information as a Reinforcer of Negative Attitudes Toward Black Americans, Elias Blake Jr., Institute for Services to Education

170 Discussion: James J. Gallagher, University of North Carolina

General Session

Educational and Social Context

Longer Education: Thinner, Broader, or Higher

FRITZ MACHLUP
Princeton University

What I am going to argue in this paper will sound crude, cruel, and perhaps untrue if the words I use in my theses are given meanings other than those I intend them to convey. Yet, if I first defined the terms, my main thesis would seem to reduce to a truism. Assuming that provocative formulations invite more interesting discussions, I take the risk of stating my basic propositions before I define the terms.

First Thesis: Higher education is too high for the average intelligence, much too high for the average interest, and vastly too high for the average patience and perseverance of the people, here and anywhere; attempts to expose from 30 to 50 percent of the people to higher education are completely useless.

Second Thesis: Longer education—education beyond high school or beyond 12 years of schooling—has become the marching order of our society; since it cannot aspire to provide higher learning, longer education can only be thinner or broader.

Third Thesis: Longer education, even if it is not higher education, may still overtax the interest, patience, and perseverance of most people; young men who have reached physical maturity resent compulsion or other pressures that impose on them several years of boredom and inactivity; the result is frustration, alienation, delinquency, and rebellion.*

---

*The proposition that the boredom and inactivity imposed on college students who are uninterested in higher or broader academic studies may lead to rebellion cannot be statistically tested, say, by correlating rebellious attitudes and academic qualifications. Students may rebel for many different causes, and some of the best students may be rebels. However, if there are large groups of students who resent the tedium of "book learning" and want to be where the action is, the probability that these groups will supply many recruits to rebellious movements will hardly be denied.

Fourth Thesis: If longer education becomes mainly thinner education, a given curriculum being stretched out over more years—for example, a 16-year program covering what can be learned in 10 years—it will have disastrous effects upon the working habits and attitudes even of those students who do not reject the system but who submit to it contentedly or in apathy.

Fifth Thesis: If longer education is broader in that it adds subjects and approaches to those taught in secondary school, it may perhaps hold the attention of the more patient ones in the age group; but we cannot expect any substantial benefits either for the graduates or for society.

It must have become evident that my definition of higher education is not the one commonly used. Those who talk about "universal higher education" or "higher education for everybody" are not talking about what I call higher education; they mean, in my terms, "longer education for everybody" or "universal post-secondary schooling."

I do not define higher education by the age of the student, or by the number of years of prior schooling, or by the name of the institution. A student who is over 18 years old, has had 12 years of school, has graduated from a high school, and is attending a college or university, is not necessarily getting what I call higher education. Not even a man 24 years old, with a bachelor's degree, registered for full-time study in a graduate school, is necessarily getting higher education in *all* his courses. He may be taking Elementary French or Intermediate German in order to prepare for a language examination—if this is still required in his graduate school. Such language study is not part of higher education, either graduate or undergraduate; it is elementary, or at best secondary, education. Likewise, an undergraduate who takes a college course in Elementary Algebra or in Trigonometry is not engaged in higher learning but rather is making up a deficiency in his secondary-school program. It is, of course, desirable or even imperative that colleges and universities offer such courses, where students can fill gaps left open in their elementary and secondary education. But, I repeat, the fact that a course is given in a university building to a student in the right age group and with the right number of years of previous schooling does not make it part of higher education.

There is nothing new about the fact that colleges and universities include elementary and secondary education in their programs; the question is merely whether the share of higher education has been diminishing. The mixture has been different at different institutions. The most prestigious colleges and universities have offered much heavier doses of higher education, heavier than would have been possible at institutions with academically less talented, less prepared, and less interested students. Even in institutions of high prestige it has usually been possible for some students to get by with a selection of subjects and courses that could not be characterized as higher education. At the other end of the scale, there have always been colleges and universities that competed for students by offering academic programs that made it easy for the academically untalented to qualify for passing grades. Thus, our colleges and universities have always been institutions of higher *and* broader education, with various mixtures of height and breadth.

This statement is not meant to refer to extension classes, to programs in continuing education, to evening schools designed for adult education. It refers to the college curriculums for regular full-time students who are willing and able to expose themselves to broader education, but who would not be willing or able to receive higher education.

I suspect that the mixture has been changing in recent years in the direction of lesser height (or depth) and greater breadth. This probably has not happened everywhere; a number of colleges and universities have been receiving better prepared students from academically strong high schools, which enabled them to step up the level of undergraduate instruction in several subjects. In most places, however, the admission of academically less prepared and less interested students has made it necessary to offer less demanding courses and to reduce or remove requirements that compelled students to include at least some higher education in their course programs. The statistical facts that in the last 20 years the percent of the age group enrolled has more than doubled, and that now more than 55 percent of high school graduates enter college, have made the dilution inevitable.

Groping for a definition of genuinely higher education, I shall approach it by way of analogical reasoning. No sane person can expect 30 percent or 50 percent of all adults to be able to run as fast or jump as high as the outstanding runners and jumpers in the country. The intellectual capacities of human beings are even more unequal than the physical, and it is patently impossible for 30 or 50 percent of the people to aspire to approach the intellectual feats that can be performed by excellent brains. I have never been able to run very fast or jump very high; I have been just about average in these physical skills. If I had taken physical education for 20 years—far beyond secondary school—I doubt that I would have become a much better runner or jumper. Similarly, I doubt that most people would become experts in higher mathematics if they took courses in mathematics for 20 or more years. All of us have limits which we cannot stretch by trying for several more years.

I define higher education as the level of scholarly teaching, learning, and researching that is accessible to only a small fraction of the people. Any level of education that is designed for a larger portion of the population is, if extended beyond the age of completing high school, in fact only continuing secondary-level education. An affluent society can offer continuing education to as many people as may want to take it. But we should not kid them, and still less ourselves, by the fake assertion that this is higher education.

I have spoken of "outstanding" athletes, of "excellent" brains, and of a "small fraction" of people qualified for higher education. Can these restrictive terms be quantified? In order to be outstanding or excellent, what is the top percentile of achievement that deserves these designations? Everybody is able to lift some weight, but how heavy a weight does one have to lift to be a weight lifter? Few of us know the answer to the last question, but we might guess that nobody would be called a weight lifter unless he could lift weights which only the strongest five percent of the people are able to lift. The highest percentile of academic achievers qualifying for higher education may also be five percent, but perhaps the economic demand for highly qualified scholars influences the standard applied. Thus, we might stick to the top five percent for weight lifters, ski racers, and concert pianists, while going to ten or fifteen percent as the fraction we regard as qualified for higher education. Some of the testing experts assembled for this conference may have an idea whether in the spectrum of academic capacity there is at some point a gap that suggests a dividing line, provided of course that capacity includes intelligence and reasoning power as well as motivation and perseverance.

Some of the classical definitions of higher education are much more restrictive than mine. Wilhelm von Humboldt, for example, held that, while the "school" had the task of disseminating received, accepted knowledge, the "university" was concerned chiefly with new knowledge; he insisted on a strict separation of higher education from schooling and on an exclusive concern with the "pure idea of scholarship." Incidentally, while Humboldt had many quite uncomplimentary things to say about the professors—he called them the most intolerant, unmanageable class of human beings—he also affirmed that professors are not there for the sake of students, but both students and professors in the university are there for the sake of scientific and scholarly research. Clearly, undergraduate education was not higher education for the man who helped establish the University of Berlin in 1809.

My own concept of higher education may be judged to be nonoperational, but it has operational variants. One operational definition could settle on admission standards as the criterion, though there is much arbitrariness in selecting the indicators of academic capacity and, as I have said, in drawing the line between the qualified and the nonqualified. Alternative definitions may use as characteristics the intellectual level required by the subjects and approaches or techniques, by the courses, laboratories, seminars, and readings. These criteria raise again the question of where to draw the line; should the subject matter taught or researched be accessible to the best 10 percent of the secondary school graduates or to the best 15 percent? Moreover, we must be aware of a continual process of downgrading some subjects and parts of subjects as they become knowledge comprehensible to young pupils of limited background. For example, differential calculus after its invention by Leibniz and Newton was surely for some time part of higher learning; but it has become part of secondary schooling—it has, in fact, been taught in secondary schools in Europe for at least a hundred years.

What I have said may sound smug, snobbish, and sanctimonious; it will be criticized as hopelessly anachronistic, as out of fashion, perhaps even as a symptom of class bias or racism. I am out of fashion, I know, but I plead not guilty to charges of class bias or racism. I firmly believe that *higher* education should be open to all who want it and can take it. But we cannot change the fact that perhaps 80 percent of the people find it "not relevant" to their interests and capacities. This is especially true of those who have been denied an adequate preparation at home and at school. *Broader*, continuing education also should be open to all who want it, and many more will be qualified for it. Moreover, I am convinced that *higher* education is not a prerequisite for political leadership or for business management, though *broader* education may be helpful. I even believe that most of those who are best qualified for higher education are not particularly suited for positions of leadership, either in politics or in business. Thus, they are not likely to govern the nation or to exercise power. Scientists spend their time in the laboratory, and scholars in the library. If I admit to being out of fashion, it is chiefly because I want undiluted higher education for scientific research and scholarly learning. What I deplore is that virtually all colleges and universities are reducing academic requirements and the level of their offerings in the name of social justice and equality of opportunity—that is, in order to accommodate more of those who are not prepared to take higher education.

It is undeniable that for hundreds of years the sons of wealthy parents have gotten into college even if they were not qualified to receive higher education; they received what they were qualified for—broader education. I am not sure that they benefited greatly from it, but they had some fun and, later in their lives, looked back on the experience with pleasure and nostalgia. Now this can be interpreted, if you will, as a "class privilege." But it would be quite unreasonable to conclude from it that the "underprivileged" should not only have the same chance but should be pressured into what to most of them is an ordeal of boredom and repression. I explain a large part of the rapid increase in college enrollment through pressures of various sorts: parental pressures, peer group and other social pressures, the hope for draft deferment, and the fear that jobs in industry and trade will be available only to college graduates. In other words, millions of young men have entered college for reasons other than an interest in academic studies. From the enormous increase in uninterested, bored, dissatisfied, and rebellious students has resulted a stampede to restructure the institutions toward further relaxation of academic standards and requirements and further dilution of the intellectual fare they provide to the students. There is a serious danger that undergraduate education will in this process sacrifice the breadth which has thus far been substituted for height or depth, and will become the endpiece of an extended string of school activities: longer and thinner education.

Ten years ago, when I wrote my book, *The Production and Distribution of Knowledge in the United States*, the trend was clearly visible; it made me speculate about the effects which the increase in college enrollment relative to population would have upon educational standards. I wrote: "Most people *can* learn what they will ever learn in school in eight years, and if they are kept there for 10, 12, 14, or 16 years, they will merely learn it more slowly" (1). I discussed the effects of prolonged schooling upon various types of students and upon society as a whole; and I concluded with proposals to strengthen preschool education, to start elementary education a year earlier, to compress secondary education, and to lower the school-leaving age.

My proposal to compress and shorten the first two levels of education is not inconsistent with the ideal of lifelong education. *Earlier* termination of compulsory, formal, full-time education is fully compatible with *no* termination of voluntary, informal, part-time education. The latter can take many forms: evening classes, midcareer leaves for one or two years of academic study, reading and discussion groups, and continual individual reading of books. In addition, my proposal included the widest possible extension of the opportunity to go to college—as institutions of broader education.

Under my plan of a more concentrated curriculum at elementary and secondary school, all or most students would "complete high school at age 14 or 15. Perhaps half of them could go on to colleges, which would receive students better prepared in English, foreign languages, and mathematics than at present, but which otherwise need not raise standards much above those maintained now [1960]." The proposed increase in the percentage of people going to college was in line with strong public demand. Going to college has become an element of "U.S. democracy," "equality of opportunity" and the "American standard of living," and it would now be politically intolerable to disappoint so many "who believe that those without college education are second-class citizens" (2). But we should not entertain the fiction that this college for the masses can offer higher education; and we should not waste the best years of our young people by pressuring them to spend 4 years in broader education on top of 12 years of previous schooling. My endorsement of as many as 50 percent of high school graduates going on to college was subject to the compression of school education.

One of my proposals, preschool education, has been put into effect, partly through Operation Headstart, partly through the rapid spread of nursery schools and kindergartens, and partly through "Sesame Street," the successful TV program. The second proposal, to start elementary school at age five instead of six, is in effect in England and may have some chance of eventual adoption in the United States. The third proposal, to strengthen the curriculum of elementary and secondary schools and compress the 12 years into 10, is probably most strongly resisted. Frankly, I am not optimistic about its adoption. The probability that colleges will receive their freshmen three years earlier than they do now is therefore not high. Still, I am not giving up hope; for what longer education for the masses involves will become more and more evident. To have 50 percent, and soon perhaps 60 percent, of our young people spend 16 years in school is economically wasteful, socially harmful, and politically explosive. It also fosters an anti-intellectual attitude in so many people that the future of American civilization may be in danger.

A clear and present danger to college education is the current rush to make offerings and requirements more relevant to the interests of the academically uninterested. There was first the noble idea of giving many more people the opportunity of an academic education. When the newly-admitted found that they did not like the traditional college education, they demanded something "more relevant." The academic departments, the committees of the faculties, their chairmen, deans, and all the rest, realized that what they had been offering was not appreciated, and they hurried to restructure the curriculum: They wanted to make it acceptable to those who do not really want *any* academic education, not higher, not broader, not thinner.

Permit me to illustrate the present trend by an analogy addressed to the music lovers in my audience. Most musicians regard chamber music as the highest form of music; there are chamber music societies in many cities and towns. Assume a movement gets going that demands "chamber music for everybody," and the chamber music societies invite everybody to join and share their pleasures. The new members will, of course, be bored, and soon they will rebel, abolish the classical string quartets, and replace them with happy rock and roll, which can "turn them on." Higher education, I submit, is to "education for everybody" what chamber music is to "music for everybody." The late string quartets by Beethoven—say, Opus 130 with the Great Fugue, Opus 133—are not "relevant" to 95 percent of adult Americans; if as many as 50 percent of the people were exposed to this glorious music, they would call for the destruction of the chamber music society.

However, our chamber music society—if I may continue to refer by this analogy to our institutions of higher education—was really not all that pure. Besides great music, they played also lighter music, to keep some of the subscribers who could not stand a program of exclusively difficult stuff. But they stayed away from folk music, jazz, and rock and roll. Today, as the new membership exerts its power, all barriers come down and the Great Fugue is drowned out by the sounds of the Rolling Stones and other attractions of the Woodstock culture. Not that I reject the new youth culture; but I grieve over the destruction of the old, esoteric culture. Its destruction is sheer vandalism, since the two cultures could peacefully coexist if only they were allowed to be kept apart from each other.

Now enough of this analogy and back to direct and blunt speech. The college, be it for the 18- to 22-year-old or for the 15- to 19-year-old, cannot reasonably be demoted officially to the rank of secondary education, even if its entire intellectual fare is to become precisely that. Those who, at no small personal sacrifice, elect to go to college would feel cheated if high school and college were lumped in the same category. But if we continue to call *all* colleges institutions of higher education, we have no designation for those institutions that are really dedicated to higher education. The difficulty disappears if we use the term "tertiary education," as many writers on education have been doing for a long time. Analysts can then, if they are interested, rate various institutions according to their "product mix"—that is, the proportions of the different levels of academic performance in the programs offered by the faculties and elected by the students.

However, terminology and taxonomy are surely not the essential problems. An urgent problem for educational policy is how much remedial education should be offered in college for academic credit. It is, unfortunately, a fact that many, many people of college age have had such poor schooling that they badly need remedial English, remedial arithmetic, remedial algebra, remedial basic skills of all sorts. Should these victims of poor schooling be admitted to college and given all the consideration that would be needed to allow them to stay in and to graduate with bachelor's degrees? Should we institute special preparatory post-secondary, pre-tertiary systems for them? Are there other options we can provide? The worst possible option, I believe, would be to admit all comers regardless of academic interest and capacity, and then either flunk them out after a year, or lower academic standards to get them through the mill and out, armed with a degree that will have lost its meaning as a certification of academic achievement. If undereducated high school graduates desire longer education to give them what they have failed to get from the school system, such education must be provided. I question, however, whether colleges should assume this function; if they do, they may lose their discernment of academic quality, which it is their prime function to maintain.

I come to my last point. Repeatedly it has been asked what percentage of the population would benefit from higher education. The question has been given different answers. The Truman Commission in 1948 said 49 percent; the Hollinshead Report of 1952 said 25 percent; and several other percentages have been suggested. I submit that the question is insufficiently specified and therefore unanswerable. It fails, first, to specify the meaning of benefit. Does it mean an increase in skills and efficiency, or in earning capacity, or in personal satisfaction? The question, secondly, fails to specify what is meant by higher education. Does it mean the academic programs of undergraduate instruction as constituted at the time, or academic programs redesigned in particular ways, or perhaps any kinds of program that appear "relevant" to the students? If we define higher education in the way I have suggested, the question becomes circular and the answer tautological, since the adjective "higher" in this definition implies accessibility to a given small fraction of the population. The answer implied in the definition, depending on its restrictiveness, would be 5 or 10 or 15 percent.

The question makes better sense if it is reformulated to ask what percentage of the population would benefit, in terms of individual increases in earning capacity from longer education, to wit, from adding four years to the given twelve-year program of primary and secondary education. We would still need some specification of the program offered or required in the additional four years. Let us assume that it refers to the type of program now being provided in typical public colleges and undergraduate divisions of public universities. It takes great boldness even to suggest an answer. The various economic studies which have shown the existence of earnings differentials attributable to college education do not throw much light on the question. First, these studies rested on data about earnings in times when the percentage of the age group going to college was about one-half what it is now; the supply of college graduates has of course much to do with their earnings, though we do not know how much. Secondly, the pecuniary benefits during the graduates' later careers yielded net returns only if all earning foregone during the years of study had been taken into account. Since these necessary sacrifices of income have been going up and are increasing from year to year, it is conceivable that future differentials in earnings are too small to yield positive returns on the investment in the students' capacities. There are still other factors complicating the problem. But I would not have the courage to predict any pecuniary net benefits from college education if more than 30 percent of the age group attend college after having completed 12 years of earlier schooling.

These considerations have been in terms of personal economic benefits to the college graduates in the form of increased earnings in their later careers. There is the possibility that the economic benefits to society as a whole exceed the sum of the private income increments earned by the graduates. Unfortunately, the opposite is also possible. It is even possible, as I explain in my recent book, *Education and Economic Growth*, that "the private rate of return on the investment in additional education . . . may be high while the social rate is zero" (3).

The question of the percentage of the population that could benefit from higher education was perhaps not intended to focus on economic benefits, but rather on psychic ones, on subjective satisfactions derived from studying and from having studied. On this question I must bow out.

REFERENCES

1. Machlup, Fritz. *The production and distribution of knowledge in the United States*. Princeton, N.J.: Princeton University Press, 1962.

2. *Ibid.*, 141.

3. Machlup, Fritz. *Education and economic growth*. Lincoln, Nebraska: University of Nebraska Press, 1970, 38.

Testing: Americans' Comfortable Panacea

THEODORE R. SIZER
Harvard University

My central thesis, very simply stated, is that America's greatest crisis rises from persistent inequities among races, ethnic groups, and social classes; that the formal education system must play a significant role in erasing these inequities; and that the testing fraternity has a significant role to play in this process. While I am aware that this analysis marks me as an old style liberal, quite out of fashion, I persist in the belief that a good society is one which, while respecting actual diversity, is open. Within the limits of their talents, individuals should be able to choose their life style and careers—to enjoy rock and roll or Beethoven quartets, as Fritz Machlup differentiates. It is the responsibility of education to make those talents as broad and deep as possible. Testing must identify and record talent, but always with a minimum of group bias. This latter task alone is a difficult one—and, as a look at the recent history of the testing movement suggests, one that too long has been slighted.

The stereotypical twentieth-century American is the engineer. A spiritual descendent of Benjamin Franklin, he is the compleat tinkerer, the man who takes someone else's theories and puts them to constructive use. He is a builder—of railroads, bridges, rockets, moon capsules, and mass education systems. His approach starts from technology—the way things work or can be made to work—rather than from pure, or speculative, science. He spends far more time and resources on developing and marketing Kleenex than on discovering the fundamental biochemistry of nasal drip. He finds ideas and concentrates on putting them to use: the internal combustion engine, pasteurization, atomic fission. And mental testing.

The American mental testing movement is largely a series of variations on the speculations and experiments of Alfred Binet, a Frenchman. At the simplest level, Binet was experimenting with techniques of sorting children. On certain supposedly status-free measures, youngsters might be separated out by mental abilities, and classified not only in terms of their current achievement, but, more importantly, in terms of their likely future competence. By the turn of the century, America was deep in the first stages of mass education and desperately needed a device for sorting children that was consistent with the movement's egalitarian ideology. Sorting by income, or accent, or conduct, while practiced, could not be publicly defended by the elected schoolman or even the appointed superintendent. Some more politic device was needed, and two emerged—and both are yet very much with us.

The first was the local control of schools, a device for a drastically decentralized school system, which wrapped strict class and ethnic segregation in a mantle of liberal political ideology. The schools must be close to the people, it was argued, and "the people" in this instance were those who lived in a limited geographical area. Control by "the people"—good egalitarian ideology—was in this instance used to defend ethnic group and class ghettoization. Americans added to this insult by then preaching that these school districts, many of which were gerrymandered enclaves, were some sort of classless melting pot. The fact that there were several prominent communities where useful mixing did in fact take place gave credence to the notion that this was the common American way.

The second device for sorting came from the clever mind of Alfred Binet: mental testing. In a democracy, Americans thought that if there should be any hierarchy at all, it should be a hierarchy of talent. Tests were needed to "prove" the existence (or nonexistence) of such talent. Not surprisingly, then, Americans engineered the idea of mental testing and adapted late nineteenth-century European theories to the realities of a more modern America. Terman, Thorndike, and the rest were pioneers, but more as engineers than as theoreticians. Terman's variations on Binet put the Frenchman's work into American terms. His writings were explanations of a method rather than expositions of the basic theoretical underpinnings of the ideas of mental testing. American experimental psychology was lively and productive, but used the basic laboratory approaches that had been accepted earlier in Europe.*

---

*To say that the American contribution to the mental testing movement was primarily in application is not necessarily to denigrate it. Engineering requires immense skill and imagination, and the translation of theory into useful devices that help people can hardly be seen as less noble than "pure" inquiry. But engineering unre-

The tests so developed were seized by schoolmen and the public to help sort people. World War I gave a massive fillip to the movement: our government had a real, and instant, need to fill round holes with round pegs and to identify potential leaders. American scholars interested in testing were drawn into these massive War Department "sorting" projects. By 1920, we were a nation that fully believed that every man had native intelligence of a certain power; that this power remained relatively constant during an individual's lifetime; and that the power could be measured, even in childhood. The "intelligence quotient" had been popularized. Democracy had a replacement for hereditary distinctions; we would be a nation with an aristocracy of God-given talents rather than an aristocracy of birth. If the mental testing movement had not emerged from Binet's laboratory, it would have had to be invented. Americans, committed politically as they were to a vague sort of egalitarianism, *needed* testing.

The movement, so popularized, quickly became distorted. Before tests were fully reliable, they were accepted as panaceas. While the scholars at the head of the movement were aware of this, public demand still ran ahead of research and development—a state of affairs, one must say, all too characteristic of American educational history. The zenith of the popularization of mental testing is distilled in a remarkable address by G. B. Cutten on the occasion of his inauguration as President of Colgate University in 1922. Cutten devoted his remarks to an analysis of "Democracy." "Let us look the question of democracy fairly in the face and be honest with ourselves," he asserted. "We are ruled in industry, in commerce, in professions, in government by an intellectual aristocracy. We have never had a true democracy, and the low level of the intelligence of the people will not permit of our having one. We can not conceive of any worse form of chaos than a real democracy in a population of an average intelligence of a little over thirteen years." He went on: "There must be some solution to the problem of government, and we must find it. What is it? We must first recognize that we are and have been, since the revolt against autocracy, ruled by the intelligentsia; more than ever the rule must be by an aristocracy, i.e., a rule by the best. . . . This aristocracy must inevitably be the most intelligent, but it must also be well

---

freshed with theoretical questioning runs the severe risk of becoming inappropriate or worse, just as theoretical inquiry unchallenged by practicality can become irrelevant. Inappropriate engineering, however, can often hurt people. Irrelevant theories rarely do.

trained, benevolently inclined, and willing to admit any others to its membership who are fitted to belong. Democracy then comes to be a government of the people, for the people, by all those of their number fitted by intellectual ability, moral ideals, and careful training. The ruling has always been by the few intelligent members of the community or the nation, and in America the aristocracy has always had the 'open shop.' The training has also been a factor, even if an accidental factor, but the element most lacking has been the moral ideals. Government for the people, instead of for the governors, must be the keynote of the future, and the task of the colleges and the universities is the training of this aristocracy.

"It may be interesting to speculate concerning the effect of mental tests upon the problem of democracy. If the present hopes and expectations are realized, they will result in a caste system as rigid as that of India, but on a rational and just basis. We are now examining children in the public schools, and find all ranges of intelligence from imbecility to genius. We are told that the intelligence quotient of a child rarely changes, so that we are enabled to tell early in his life what the limit of intelligence of any person will be, and in a general way to what class of vocation he is best fitted, and, to a certain extent, destined. When the tests for vocational guidance are completed and developed, each boy and girl in school will be assigned to the vocation for which he is fitted, and, presuming that the tests are really efficient, he will in the future not attempt any work too advanced for his ability and hence make a failure of it, neither will he be found in an occupation too elementary for his ability and hence be dissatisfied. Economically nothing could be more desirable. All differences in accomplishments or results from that which the intelligence quotients would indicate would be due to certain traits of character which intelligence tests do not measure, viz.: industry, perseverance, thoroughness, honesty.

"One's intelligence quotient will eventually be known and persons will be classed thereby. Those of high intelligence will be directed into lines of occupation which call for leadership. Those persons will naturally be placed in the professions, and in leading positions in industry, commerce, and politics. Each person will then be directed on a scale of intelligence down to those whose work is of the most routine character of which an imbecile is capable. But what effect will this have on our so-called democracy? It must inevitably destroy universal adult suffrage, by cutting off at least 25 percent of the adults, those whose intelligence is so low as to be incapable of comprehending the significance of a ballot. On the other hand, it will throw the burden and responsibility of government where it belongs, on those of high intelligence, and we come back again to the rule of the aristocracy—this time the real and total aristocracy. For its own salvation the state must assume the obligation and responsibility of selecting this intellectual aristocracy, and having selected it see that it is properly trained" (1).

Such was the optimism about mental testing in the Age of Warren Harding—and before George Orwell and Michael Young. There were balloon prickers even then, of course, and none more caustic than Walter Lippmann. Writing a series of articles on testing in the *New Republic* during 1922, Lippmann made much of the lack of congruity between the Terman-Stanford-Binet formulas and those that emerged from Yerkes' work with Army recruits during World War I. He was scornful of the gullibility of many concerning the general validity of supposedly standardized measures that had evolved from tests of very small numbers of very homogenous children. "The real promise and the value of the investigation which Binet started," Lippmann wrote, "is in danger of gross perversion by *muddleheaded and dangerous men*" (2).

Lippmann, however, was a cranky exception. Optimistic Americans preferred to believe in the existence of measurable, innate intelligence. To this day, mothers weep over the results of I.Q. tests—and teachers assign their children to categories with the arbitrariness of medieval jailers. And to this day, Americans like to believe that while all of us are created equal, some are more "intelligent" than others, and measurably so. Americans avoid defining this condition they call intelligence; but the layman still believes, as did President Cutten of Colgate, that it is real and fixed. And if some groups appear from tests to be less "intelligent," too many of us still say that they, alas, are congenitally stupid.

In sum, Americans needed mental testing to help classify children in school. They rushed this process into use before sophisticated and broad-scale research could properly be completed. In their haste for a system to cope with the large numbers of children, they brushed aside some glaring inconsistencies—the lack of congruence between Terman's and Yerkes' findings, for example—and supported the system with rhetorical hyperbole. If one is in a cynical mood, one can further speculate that many Americans quietly applauded the finding that proportionally more children from well-to-do families scored well—that is, were considered of innately higher intelligence—than those from low income families. It reinforced the smug belief that those running the country were, in fact, by natural selection the most intelligent. Mental testing for all produced a classification that roughly followed class lines. Local control further provided for safe class and racial enclaves. Together these two pillars of egalitarian idealism—mental testing and localism—largely guaranteed antiseptic and segregated classrooms for the upper middle class. Even the Lynds' study in the mid-1920s of "Middletown," a community of varied social classes but with a single "melting pot" high school, revealed the actuality of class segregated education. The youngsters were classified by supposedly "objective" mental tests—and ended largely in socially homogenous classroom enclaves. To be fair, one cannot suggest there was—or is—a conspiracy to use tests to keep poor people down. One can ask, however, why scholars in the measurement field were incapable of preventing the distortion of their ideas as these were popularized and put to use. Three explanations are plausible. First, and most obvious, is that so many of the leading scholars in the field were involved principally in the engineering aspects of it—developing minimally satisfactory tests for use by the schoolmen who were frantically demanding them—that they lost sight of the forest for the trees. One need only look at our own pell-mell rush in the last decade to computer-aided instruction to see how easily perspective is lost among the ablest of men.

A second reason is equally obvious: the country found mental tests so compatible with its ideals and its practices that it deliberately closed its collective ears to the counsel of scholars. Even Lippmann caused only a ripple. His reasonable critique, ironically, stung the experts more than it educated the general public. Perhaps their sensitivity is a measure of the misgivings that they preferred to keep smothered. The tests did reinforce class bias. But America did not want to lose its faith in a system that filled a needed role so satisfactorily.

A third, and less obvious, reason may be found in the narrow outlook of the leaders of the American movement. Mental testing in the United States came substantially out of the traditions of experimental psychology and statistical measurement. The laboratory approach was nonhistorical; it called for a careful study of a few phenomena at a particular time. Great attention was paid to the subtleties of activities of a relatively few subjects, and statistical analysis was expected. As a result, most research involved few subjects, most often drawn from a narrow social class group. Significantly, the only broad survey undertaken, the Armed Forces study, was headed by a psychobiologist, Robert Yerkes, a man notably skeptical of the statistical approach to mental testing and a critic of Thorndike.

The laboratory approach taken by Thorndike and others was not necessarily unwise; it was, rather, incomplete. The *developmental* aspects of intelligence were slighted. Insufficient attention was given to how "intelligence" (however defined and measured) appeared to change over time, and what caused this change. Sociological and anthropological issues were neglected, and fundamental issues such as the effects of heredity and environment on intelligence either ignored or sloppily treated.

Save for Yerkes—who, after arriving at Yale in 1924, spent most of his research effort on animal psychology—the mental testing field was dominated by a group of scholars who had similar training. A striking number were trained at Columbia by James Cattell—Thorndike, Woodworth, Kelley, Dearborn, and others. This group served as an "invisible college" and dominated the field because, for the consumer, there were no alternatives. The questions of environmental influence, of the effects of social class, ethnicity and race, and of developmental patterns of change on an individual's measured intelligence have had to wait for a group of scholars trained in more diverse ways and sensitive to a broader social experience than were Thorndike and Terman.

Americans used mental testing to give their schools, all too often, more the appearance than the substance of democracy. Characteristically, they oversold the virtues of this convenient, scientific, egalitarian system; they liked what it produced and so marketed it with enthusiasm. Again characteristically, the experts in the field were either unable or unwilling to check or moderate its popularization. And so the mental test became a well-established educational panacea.

How can we use it today? As 70 years ago, we need a device that can democratically predict the achievement of children in a variety of skills, some intellectual, others vocational and affective. Above all, we need a system that both accommodates the effects of the environment and points the way to lessening its effect.

It is no longer hyperbole to note that this country is in the midst of a social revolution, and perhaps on the brink of a violent one. The facts of the matter can be boldly stated: one-twentieth of Americans control one-fifth of the wealth and one-fifth of Americans make do on one-twentieth of the wealth—and this reality has not changed, relatively, since 1945. Both liberal and conservative, if with different rhetoric, now applaud segregation; we are once again hearing justifications for a nation of enclaves from liberal spokesmen. They properly urge cultural diversity, but they fail to face the fact that freedom within diversity requires *understanding* and *toleration* among groups. Equally important, it implies *openness* among groups; a free, if diverse culture must allow individuals to move from group to group. Enclaves may give us diversity, but enclaves without open doors will stifle freedom. The educational system carries a special burden both of encouraging those attitudes of tolerance and justice among youngsters and of teaching skills that allow individuals to move from one group to another. We need tests to show the *development* of an individual's capacities and attitudes, tests that carry as little class bias as possible. Tests of varied qualities must be developed. We need not only "intelligence quotients," but also "bigotry quotients"—and remedial work for youngsters who are excessively bigoted. I am not being facetious here; the moral development of a youngster—his sense of justice and his use of justice—is perhaps more important than his cognitive development. This country has suffered excessively already from intellectually able, but morally stunted people.

Put in a different way, the testing fraternity needs to concentrate on the effects of class, race, and ethnicity on the development of skills and attitudes. It needs to help us understand how these factors influence human development over time. It needs to suggest ways of lessening those influences that narrow a youngster's options, and ways of measuring the child's progress in increasing his options.

Testing must not in a benign way serve as a device to preserve the social status quo. On the contrary, it must be used to illumine current social rigidities—and to help us finally break out of them.

REFERENCES

1. Cutten, G. B. The reconstruction of democracy. *School and Society*, October 28, 1922, XVI, 409, 479–481.

2. Lippmann, Walter. The mental age of Americans. *New Republic*, October 25, 1922, 412, 215.

Social and Cultural Change and the Need for Educational Information: The Futurist's View

HERMAN KAHN
Hudson Institute

When I was first asked to give this talk, we had been negotiating with the Department of Health, Education and Welfare about doing some studies on the educational system. I had assumed that by the time I gave this talk I would know a good deal about the system and the main issues.

We did not get that contract so we did a little work on our own, and now I am hopelessly confused about the educational system and educational tests.

I am going to try to focus my attention on some matters that I think will be of general interest and make a few comments on how they might relate to the testing issues. Let me start by emphasizing a certain point of view and then, in effect, try to knock it down.

It is useful, often, if you are trying to impress people, to use very large numbers. You may refer in passing to the fact that there are a hundred billion stars in a galaxy. How many people count that high? Very few. Or refer to the fact that man has been on earth for about two million years? Very few studies, you know, go back that far.

Let me use this technique: Man has been on earth for about two million years. I have studied every one of those years rather carefully and I have found only two incidents of any interest—the rest were basically trivia. An unbelievable amount of trivia.

The two incidents of interest were the agricultural revolution and the industrial revolution. Now, I must concede that if you are a religious individual, you have to add a third incident, but we might disagree as to what that incident is. My personal choice would have to be the covenant of God with Abraham. But views vary. Let us therefore leave this out and concentrate on the two noncontroversial incidents.

The first interesting incident, the agricultural revolution, created civilization. For every 20 people on the farm you had a man in the city—and therefore civilization—for civilization means civic culture, living in cities. We all know what it is: bureaucracy, taxes, armies, classes, educational testing, and so on.

It did not increase the per capita income as far as we can tell. Basically, man has lived—and this is a very misleading remark, but useful—on something between $50 and $250 per capita throughout history until about 200 years ago.

The next revolution, the industrial revolution—which conventionally we say began about 200 years ago—changed per capita income. The British learned the trick of enlarging their population by 2 percent a year and their productivity by 1 or 2 percent a year per capita, and that was the wonder of the age.

We now believe that the next 10, 20, or 30 years will see as great a revolution in mankind's history as the first two. We call it the achievement of post-industrial culture, and it needs to be seen in relation to the first two. For example, industrial culture is the post-agricultural culture, if you will. Instead of having 95 percent of your people doing farming or fishing or mining or forestry, in the United States today about 3 percent of the people manage to do almost all these things for us.

The basic goals of all our industrial effort might be some such notion as the Chinese concept of the five guarantees: adequate food, adequate clothing, adequate shelter, adequate medical attention, and adequate funeral expenses. One might argue that the post-industrial culture will have no interest in these guarantees, but that would be putting it too strongly. You really could not say that agriculture is unimportant in the United States today: It is just unexciting. With only 3 percent of the people involved, we are tempted to just watch and say, "Thank you." And generally we forget to say, "Thank you."

So the success of farming has made it seem dull. And the success of industry will make it seem dull. What happens after that? We do not really know. My own picture is of something like third century Greece, or, more hopefully, fifth century Greece, or most likely something between the two and at the same time very different. In Figures 1, 2, and 3, I have attempted to project some value changes relevant to this future. In Figure 2, for example, I list value systems associated with the post-industrial culture. I believe they will all come in and are, in fact, presently emerging.

The word "post-industrial," may I say, provides a good description.

Professor Machlup coined the phrases "the knowledge society" and the "knowledge industries." "Post-industrial" was coined by Daniel Bell. I myself used to use the terms "post-American" or "post-mass consumption," but actually "post-industrial" is better because it is literal. It says industry is no longer central, but does not presuppose what follows it.

Now, when you have an event which occurs in a short period of time—10, 20, or 30 years—and claim that it is as important as the agricultural revolution or the industrial revolution, you must assume that big changes will take place in everything else. Yet I want to argue, seemingly contradictorily, that the current ferment in America is, in some ways, less due to changed challenges of this sort and more due to an erosion of the old system.

If this is true, it is very important, because it could mean that the current ferment points in the wrong direction. In other words, if all this ferment is an attempt to cope with the new, then presumably it will have something to do with the new culture and society that emerges. But then, again, it may just be that the old is disappearing, and in this case all the ferment may be directionless.

Let me define a term called "middle America." We used to use the term "lower middle class," but that sounds invidious. Sometimes it is "the forgotten man," if you will. "Middle America" can be defined operationally by reference to Figure 1.

Now, if you are largely preoccupied with the issues in Figure 1, you are a member of middle America, for my purposes. These values are associated with income. That is, an overwhelmingly high percent of the people in the United States who make between $5,000 and $10,000 a year in the North (or between $3,000 and $8,000 in the South) would be preoccupied with these issues. In that sense, these issues are associated with economic classes.

But every Texas millionaire I have ever met is largely preoccupied with these issues, and an increasingly large number of people we used to call upper middle class Americans. That is, while I mainly want to talk about the erosion of these values, I actually believe that in the last 10 years there has been an increased number of people concerned with these values. This is a kind of counterreformation, if you will. It has come in part as a reaction—and maybe it is an overreaction—to the antics of the kids in prestige schools and to other events. Figure 1 probably describes a higher percent of America than it did 10 years ago.

Figure 1

THE TWELVE TRADITIONAL SOCIETAL "LEVERS"

(That is, traditional sources of "reality testing," social integration, and/or meaning and purpose)

1. Religion, tradition, and/or authority
2. Biology and physics (for example, pressures and stresses of the physical environment, the more tragic aspects of the human condition, and so on)
3. Defense of frontiers (territoriality)
4. Earning a living (for example, the five guarantees)
5. Defense of vital strategic and economic interests
6. Defense of vital political, moral, and morale interests
7. The "martial" virtues such as duty, patriotism, honor, heroism, glory, courage, and so on
8. The manly emphasis—in adolescence: team sports, heroic figures, aggressive and competitive activities, rebellion against "female roles"; in adulthood: playing an adult male role (similarly, a womanly emphasis)
9. The "Puritan ethic" (deferred gratification, work-orientation, achievement-orientation, advancement-orientation, sublimation of sexual desires, and so on)
10. A high degree (perhaps almost total) of loyalty, commitment, and/or identification with nation, state, city, clan, village, extended family, secret society, and/or other large grouping
11. Other sublimation and/or repression of sexual, aggressive, aesthetic, and/or "other instincts"
12. Other "irrational" and/or restricting taboos, rituals, totems, myths, customs, and charismas

It is very important to understand this. America is not moving for the moment to the left, but to the right. This is quite clear today.

It is also interesting to note that this movement, and even the characteristics of the people of Figure 1, were largely misunderstood by upper middle class America, or by the progressive middle class, intellectual, articulate America.

I sometimes list eight issues that were, except for the Vietnamese war, major issues in the 1968 campaign and certainly the 1970 campaign—and race and ethnicity are not among them, by the way. These eight issues were almost completely misunderstood by the scholarly community, and by sociologists and the articulate press.

I had a girl spend two weeks in the New York Public Library and in the Columbia University Library in October of 1968, making a quick check on this statement for me. She could only find some six or seven papers that presented a reasonable discussion of these issues. Let me mention two of them.

The first is the issue centering around the term "law and order," as used in the '68 campaign. As far as I can tell, we were told by every respectable newspaper and almost every respectable scholar that the term was a code word for "anti-Negro." This is peculiar, because in many American cities two-thirds of the Negros put law and order as the top issue of the campaign. And *they* cannot be anti-Negro, at least in exactly that fashion.

The term was never used by George Wallace in that way in his campaigns in the North. Now Wallace is a racist, and if you asked him about Negroes, he often told you he did not like them. He did not need code words. He was very careful to differentiate the law and order issue from the race issue. And, indeed, his focus was not solely on Negro crime or Negro race riots, but more on just crime in the streets and among young kids. These were the issues that really bothered these people.

Let me mention another issue, which to me was even more interesting: the so-called "backlash." I believe what this concept is supposed to mean is something like the following: that lower middle class Americans—or better, traditional Americans—tend to be racist in their attitudes. That is a completely correct statement. But the idea of backlash contemplates something more: that this racism, this attitude of keeping Negroes down, is increasing in America as a result of the rapid advancement, or at least the pressures for rapid advancement, of Negroes in America.

This last statement seems to me to be almost completely untrue. We have been looking at every relevant poll we could find, and we cannot find a single one that does not demonstrate—or tend to demonstrate—that racism among middle Americans is on a very rapid decline. For example, a Gallup poll in 1965 found that 64 percent of all white southerners said they would not live in a neighborhood with Negroes or send their children to a school with any Negroes. Only 16 percent take that position today.

Another poll asked, "Would you vote for a Negro for President of the United States if he were otherwise qualified?" In 1963 only 47 percent of Americans said they would, but 67 percent say they would today.

I do not know of any contradictions to these polls, but what do you see in the reports? *Newsweek* has some of the best data in the world on this particular issue. They had a whole issue on Negroes and talked about increasing racism and backlash—even though their own data contradicted this conclusion.

Time magazine had a few pages on Negroes recently. They talked about increasing backlash, though again their own data contradicted it. And so on, in the scholarly community.

This is no small oversight. For example, the only place I could find a certain Wallace meeting accurately reported was in a newspaper called the *East-Village Other*. Some of you may know it. It is anarchist, hippie, new left, protest. But for some reason or other the reporter could listen—could actually *hear* what was said. He hated Wallace, by the way, and with an intensity much greater than the respectable press, but he could listen. And he could report what he actually heard and not the results of some fevered imagination, which is really an extraordinary accomplishment these days.

This brings me to the second point I would like to raise. Figure 1 shows a series of what I call societal levers. We might call the express purpose of the lever the ostensible or manifest function. Each also has a latent function—to use the jargon, if you will—which is to keep you in touch with reality, as the subtitle indicates.

For most people, their only contact with reality is the requirement of earning a living. Take that away, and most people can, and will, live in illusion. A book called *Iron Mountain* recently claimed that a group of American sociologists under contract with the government produced a study that said the only way to maintain social unity and contact with reality was to have a need to defend frontiers. Take that

Figure 2

THE CURRENT "TRANSITION" AND/OR SEARCH FOR MEANING AND PURPOSE SEEMS LIKELY TO ENCOURAGE THE FOLLOWING:

1. High consumption, materialism, and other pursuit of middle class sensate values
2. Neo-cynicism
3. Being a human being (neo-epicureanism, familial and altruistic motivations, and/or emphasis on interpersonal interactions)
4. Fulfilling a sense of responsibility (neo-stoicism)
5. Neo-gentlemen (for example, neo-Athenians and/or Europeanization of the U.S.)
6. Self-actualization
7. Special projects or programs that create general or specific esprit, elan, pride, excitement, charisma, and/or chauvinism
8. Humanist left, responsible center, conservationists

– – – – – – –

9. Semipermanent adolescence
10. "Bread and circuses" (including, for example, both welfare and "happenings")
11. Rise of new and old cults
12. Fanatic reformism (for example, propaganda by the deed, protest by terror, violent conspiracies, insistence on immediate solutions–the "now" generation–and so on)
13. Protest, revolution, and violence as a kick or even a way of life (for example, a commitment to nihilism, anarchism, and/or neo-fascism, as well as "ordinary" protest movements, demonstrations, and riots)
14. "Drugs and fornication"
15. Other kinds of "dropouts" and quasi-dropouts
16. Emotional and "reactionary" backlashes–traditionalists

NOTE: The term "neo" implies a modern version of what occurred in third century Greece.

away, according to this report, and you could not keep in touch with reality or have social cohesion.

I would argue that any one of these 12 things can keep you in touch with reality, each in its own way. Take away all 12, with no replacements, and you are out of touch with reality, perhaps in much the same way that sensory deprivation experiments also result in a loss of reality testing. That is, if you put a man in an insulated chamber and remove all stimulation of his senses, he starts to hallucinate and/or become paranoiac.

I would say that the biggest thing going on in America is "value deprivation," and that this is the source of a good deal of our current ferment.

Now, I would like to describe some of these value changes in terms of Figure 3, because I think this may set forth the issues dramatically and reasonably. In Figure 3, I list five kinds of character structures, or *Weltanschauungs*, which seem to exist in Western culture. Presumably, for the Japanese or the Indians or the Chinese a different list would be needed.

Now these world views are not necessarily contradictory. You can find individuals who have elements of two of these, or three, or four, or all five. So in this sense it is a question of mixture, if you will. In each case, I would argue, there is a reasonable emphasis or form to each set of values, and counterposed to it there is a pathological form. Obviously this is a question of degree, and one does not want to use two-point scales, but they are simple and convenient.

I had a grandfather who used to live in Column 5, which I call "God's will." He got up every morning, talked to God, got his instructions, carried them out, and at the end of the day reported on what he had done, and checked to be sure that everything was in order.

When I was young, I put him in the bottom half of that column, down there with "bigotry" and "fanaticism." But about the time I reached the age of 30, I moved him up into the "revealed truth" cluster. It takes some of us a long time to learn.

He was a very poor man, in the sense that he lived on about 5 to 10 percent of my income, but he did not know that he was poor. He had the impression that he was rich—and certainly he lacked the confusion and identity problems so common today.

Most people attending this conference were raised in Columns 3 and 4. Now Column 4, "Conscience," is getting a bad name in

Figure 3

SOCIAL AND POLITICAL IDEOLOGIES MAY EMPHASIZE:

(1) *"Transcendence"*	(2) *"Impulse"*	(3) *"Reason"*	(4) *"Conscience"*	(5) *"God's Will"*
(Leading to, at best, a reasonable emphasis on)				
SPIRITUALITY	FREEDOM	RATIONALITY	LOYALTY	REVEALED TRUTH
PERSPECTIVE	SPONTANEITY	MODERATION	DEDICATION	DIGNITY
PAN-HUMANISM	CREATIVITY	THOUGHTFULNESS	TRADITION	SALVATION
IDEALISM	PERCEPTIVENESS	MELIORISM	ORGANIZATION	RIGHTEOUSNESS
ALTRUISM	PARTICIPATION	FLEXIBILITY	ORDER	ESCHATOLOGY
MYSTICISM	SENSORY AWARENESS	CALCULATION	OBEDIENCE	WORSHIP
DETACHMENT	SELF-ACTUALIZATION	PLANNING	SELF-SACRIFICE	AWE
REVERENCE	JOY AND LOVE	PRUDENCE	JUSTICE	SUBMISSION
(But with a corresponding potential for a pathological degree of)				
DROPPING OUT	PERMISSIVENESS	ABSTRACTION	AUTHORITARIANISM	BIGOTRY
PASSIVITY	IMPULSIVENESS	THEORY	RIGIDITY	FANATICISM
MYSTICISM	ANARCHY	RATIONALISM	RIGHTEOUSNESS	RIGHTEOUSNESS
CULTISM	LAWLESSNESS	INDECISION	DESPOTISM	DOGMATISM
UNWORLDLINESS	CHAOS	DEHUMANIZATION	SADOMASOCHISM	HYPOCRISY
SUPERSTITION	NIHILISM	SCIENTISM	PUNITIVENESS	SUPERSTITION
WITHDRAWAL	VIOLENCE	DOGMATISM	FANATICISM	PASSIVITY

America today—it has a connotation of neurotic guilt, rigidity, or just being generally "hung up." We would say that the Nazis belong in the bottom half of that column and ourselves in the top half. The young kids at the prestige schools would put us in the bottom half of that column, along with the Nazis—and I do not know who they would put in the top half.

As you know, one of the things about the young today that bothers many people in this room more than anything else is their animosity towards reason, towards rationality. And the reason is clear: The young today tend to see most people who indulge in reason as belonging in the bottom half of this column: abstract, theoretical, rationalistic, indecisive, dehumanized, scientistic, and so on—and this bothers those of us who pride ourselves on our rationality. I assume, of course, that everybody at this conference is in the top half of that column.

As far as the kids themselves are concerned—and I'm talking here about the hippies, the protest groups, the new left groups, all different, but for this purpose they can be seen as the same. They try to put themselves in the top half of Columns 1 and 2. I would put 95 percent of them in the bottom half—though there are some who do belong in the top half.

Now, if you look at the top half of Column 2, you will see the values of childhood. They are very attractive in a five-year-old, but you may or may not like them in a thirty-five-year-old. The hippie and the new left, in particular, believe that a conscious attempt must be made by society to preserve these values of childhood into adulthood, and I certainly am not against that attempt. But I am against the attempt to force it on everybody. There is an enormous difference between a group choosing this as a way of life for themselves, and trying to force it on the rest of the country.

Now, it is very important to notice that the new left, the hippies, and many of the protest groups are in Column 1 also. There is a religious element there, without being religious. It is like the Unitarian Church: There is *at most* one God—though some of the kids would challenge that.

They have a religious sense, but not a religious faith, which is fairly impressive. One thing people like myself often object to is the fact that they want to tear down the old structures, but have designed no new structures to replace them. They want to destroy everything that is old—destroy chamber music without necessarily suggesting rock as a substitute.

That is really not as bad as it sounds. Hippies, in particular, have a very conscious sense of being John the Baptists. And there is no point in asking John the Baptist what the message is. He hasn't got it. He is not Jesus Christ. All he knows is that there is a message on the way. And he is trying to get the crud out of your ears, you know, so you can hear it.

I think that is not a bad position because, in a way, I share it. We have a very different concept of what Jesus Christ is going to look like, but the basic feeling is the same.

My time is about up, and I must still attempt to relate all this to education. To me the really important thing about an educational system is that it not produce people who are excessively characterized by what Veblen called "trained incapacity." I would call this "*educated* incapacity," because the problem is broader than Veblen, in my view. Veblen was referring specifically to engineers and sociologists with Ph.D.'s who could not understand large issues because of their training. I would like to extend this idea because the problem is incredibly more widespread than that, and socially very dangerous.

I have spent a lot of time with pollsters recently, and, you know, they really cannot formulate the right questions because of their upper middle class bias. They just do not understand the issues. The previous speaker referred to the fact that America must be moral. It has always been known as a moral country; de Tocqueville observed this. The speaker suggested that the definition of morality is attention to injustice, particularly racial and ethnic. That is certainly part of the definition, but it is very hard for the upper middle class Ph.D. to understand that it is only part of it.

People talk about how the school systems distort the Negro by forcing the middle class values upon him. I would like to make a stronger statement: To the extent that our school systems are designed by the present leadership of the progressive middle class, to which most of us at this conference belong, these systems do not serve the needs, or fulfill the desires, or even begin to understand the average American of any color.

In line with my earlier remarks, issues which could be accurately described by 60 percent of America were not understood by scholars and the elite press up to about mid-1969. Now, if such a high percentage of prevailing "elites" cannot understand highly emotional issues, I submit we are in deep trouble. We gave a series of briefings to congressmen of both parties and their staffs in October 1968, and their reaction to what we said was very interesting. Many of them had come out of the lower middle class, and invariably they said something like the following: "You know, you told us nothing new; we just had forgotten."

Unlike the scholar, the politician is not allowed to forget. He has to keep in touch or he loses his job. What I am suggesting is that there have to be more effective ways of using our school systems to force a certain degree of reality testing, gaining information about their own society, upon the people involved in this system—particularly those who try to design, shape, evaluate, and run it.

Let me finish with a comment on the first speaker's position, which I am largely sympathetic with but not necessarily in exactly the same way—that is we may differ on certain measures. The term "higher education" is the problem. One of the great strengths of America has been that it allowed the Edisons and the Henry Fords as well as the Steinmetzes and the Einsteins to have an impact on our culture. But the nature and value of the Edisons and the Henry Fords today goes largely unrecognized by our educational system. A different kind of education is required by this special group. And it is certainly not "lower" education, in the sense of performance or contribution to society.

I would argue that one of the great virtues of America, and Japan, and the Soviet Union, is that all of them possess very large college educated groups of people who have an ability to run an industrial society. This is largely lacking in Europe. Europe's more elitist education has failed to produce this middle level managerial group, and the lack shows up in productivity and practically everything else tied to productivity.

The kind of education needed is not higher education according to Machlup. It is simply broader and longer, I suppose. But I would object to entirely denying it the meaning of "higher" in life and in education.

Discussion

AMITAI ETZIONI
Columbia University and
The Center for Policy Research

For a conference of people interested in measurements, the three papers I am to discuss have one thing in common: They are fairly innocent of data.

The paper by the first speaker has a key concept, higher education. At one point it was defined tautologically as that education which the upper five percent or so get. In that case, there can be no quarrel. It's true by definition. There remains only the question: Why was it defined in that elitist way?

If the definition has any empirical relevance—that is, it is suggested as a proposition that 50 percent or more of the population are incapable of benefiting from the education given now to the 5 percent—there is no evidence presented here that this is the case. Maybe by the time this paper is published in the *Proceedings*, we will be able to get some documentation.

Pointing to the fact that students are rebelling, as an indication of their boredom and incapacity to absorb chamber music, seems questionable; I would suggest that this is a very poor correlation. Most students, even in my university, are studying. Given time, I could present evidence there is great benefit on all dimensions; and I believe most students are not pressured into studying, nor are they rebelling.

Actually, the students—and the minority students especially—despite the headlines in the press, are working rather hard. If you go, for instance, to the Federal City College in Washington, which is 96 percent black, you will find they are asking for tougher studies, more education, and are staying away from the Mickey Mouse courses and the easy progress.

If I may continue in this vein, should Dr. Sizer revise his paper for the *Proceedings*, maybe he can give us some evidence on what I find the most exciting point in his paper: the idea that we should have a test for moral growth, for tolerance, for our capacity to overcome bias. This is very welcome and is central to his idea that tests should be used to move society forward rather than to cement the divisions of yesterday.

We have not heard yet during this conference that such a test can be devised. Maybe it can. The more we are enlightened on this subject, the more progress will be made.

Dr. Kahn gave away the way he does research—by sending a girl to the Columbia Library for two weeks. And the information reflects the method. Of course, that was an unfair comment, but it gets at a more serious issue—the fact that the data that was presented does actually conflict with many studies I'm well acquainted with. There is a backlash in this country. And it is serious. The polls show it. It's only as it was here defined that they don't register it.

As Mr. Kahn pointed out later, the pollsters tend to ask white middle class questions. If you ask Americans, "Would you mind if a Negro would become President?" the answers seem to indicate a decline in prejudice. But you get a different answer if you ask, "Are Negroes moving too fast? Should they get fewer economic benefits? Are they getting too many of the educational resources?" If you ask this kind of *reallocation* question, which is at the center of the backlash, then you'd get a clear indication of prejudice.

I myself did a study for the Office of Economic Opportunity on the subject. I found that while the majority of Americans still favor social programs—Medicare, expanded social security—they do so only so long as they are not specially geared to the black minorities or to the disadvantaged. They make a very clear distinction. More than 50 percent feel, "The blacks already got more than they are entitled to. They are progressing too fast." And I believe that belongs in the picture.

Now, to turn to the future, the context for much of this discussion lies in the question: What is the future going to be like?

I believe none of us has yet found a technique for divining the future. Reports on the year 2000 predict an environment that is too far away to allow us any real sense of the validity of these predictions. And, predictions are often made not only in an interval, which is, of course, necessary, but in such an open-ended fashion that they are not subject to testing, even at the year 2000.

So, for instance, if we talk about increased permissiveness, we would have to expend the whole conference simply gaining an understanding of what we mean by that concept. Do we mean consistent permissiveness? Ultra-permissiveness? This year's Spock edition, or the first edition?

We need some kind of specificity before such projections can become a useful guidance to analysis and thought. Take the question of whether, and to what extent, society is moving away from emphasis on efficiency, instrumentation, bureaucracy, and rationality to a greater concern with humanism, with relations among people, with productive leisure. If for a moment we play with the hypothesis that this might be one possible direction of the post-industrial society, instead of the discussion we had here this morning, we become involved with a wholly new context. We find ourselves asking, "Why should those people be on the assembly lines producing more cars? Why shouldn't they be in college? Shouldn't we have 100 percent of the population sharing this privilege? What else is there for them to do that is so urgent?" Maybe in the society of the future, working will cease to be the central activity, and studying and community life, in one form or another, will become the central activity.

Is it such a horror to foresee a society in which our present affluence could be produced with two hours' work a day, and the rest be spent in roughly what our students are doing today?

I'm not talking about the few; I'm talking about the majority, which switches between educating itself and being educated—longer, broader, and higher—and sharing in public activities, aimed at making the society more just and humane.

Session A

Educational Applications

School Testing to Test the Schools

RICHARD M. JAEGER
U. S. Office of Education

Last year, more than 30 million children in the nation's elementary and secondary schools spent more than 50 million hours of classtime completing standardized tests, at a cost in excess of a quarter of a billion dollars. I submit that because of a failure to use the resulting data much of this time and money was wasted. Thus, I would like to suggest some ways in which test results can be used more effectively in managing the nation's schools.

For the sake of simplicity, this paper presumes a school system seeking to develop a uniform set of abilities among its students. However, the concepts presented can be readily adapted to the pluralistic goals so eloquently advocated by Edgar Friedenberg (5) at this Conference last year, and by Peter Schrag (9) in the *Saturday Review* of September 19, 1970.

In some early letters arranging this Conference, Gene Glass defined an "educational information system" as a collection of data on the behaviors of children and adults that is maintained for the primary purpose of educational decision making. By that definition, school testing programs are one component of several educational information systems. Decision making of one sort or another is almost always included in discussions of the functions of school testing programs. Though the decisions most often mentioned involve guidance and placement of individual pupils, I would like to explore the uses of test results in making decisions affecting institutions. Decisions about pupils are usually made by teachers and guidance counselors. However, institutional decisions about how school systems are managed are made by superintendents and administrators.

Several of our colleagues have suggested that school testing programs might provide information useful for educational management. In a paper appearing in the 1951 edition of *Educational Management* (11), Ralph Tyler proposed that test results would be useful in the development of policy and plans for educational programs, for the evaluation of educational programs, and for interpreting the work and needs of the schools to the public. Some of these notions predate Dr. Tyler's paper by at least three decades. In the 1918 *NSSE Yearbook*, Haggerty (7) reported the results of a survey on the uses of test results in school planning. He stated that "31 percent [of the sampled school districts] reported conscious use of test results in changing school organization, courses, pupil assignment, instructional methods, etc."

Despite this long-standing acknowledgment of the usefulness of school testing programs for educational management, there is little evidence that today's school administrators consider test results when making major management decisions. In fact, there is considerable evidence to the contrary. During the past year I had occasion to correspond with directors of research in most of the large-city school systems in the nation. I discussed with many of them the uses of institutional test results in their cities. Typically, the efforts of the research directors to promote the use of test results in educational decision making were met with resistance or apathy. The annual test reports now published by most large school districts often result in headlines in the local papers and statements of outrage by school board members for a day or two. Then they are filed and forgotten until the following year. Surely greater use should be made of information obtained through the expenditure of so much time and scarce funds.

Let us now consider some ways in which institutional test data might be useful for education management, some paradigms for the use of test results for these purposes, and some areas where the state of the testing art needs improvement if test results are to be used for decision making.

Some Uses of Institutional Test Results in Education Management

ALLOCATION OF RESOURCES

Central administrative officers in school systems make decisions on the distribution among schools of human and capital resources, such as assignment of teachers, allocation of instructional materials, and allocation of building funds. Several studies, such as those of Sexton (10) and Guthrie (6), have revealed that the distribution of resources among schools is far from uniform, and in many cases regressive with respect to educational need. In considering the use of test results as a basis for resource allocation, socio-political influences will be ignored. That is, I shall assume a decision maker who wants to allocate resources in accordance with educational needs.

MODIFICATION OF EDUCATIONAL PROGRAMS

Institutional test statistics might also be used by education managers to appraise the success of specific instructional programs. Such appraisals would be used to formulate decisions on program modification. This evaluative use of test results is common in local school systems that operate federally supported projects. In those cases, evaluation is usually required by law. However, institutional test results rarely influence the modification of regular programs of instruction, which is possibly a telling indictment of the utility of testing programs.

PUBLIC UNDERSTANDING

The test reports published by city school systems reflect a widespread concern for "public information" on the status of achievement in schools and school systems. Mean or median achievements of pupils in selected grades throughout the school system are almost always reported. Increasingly, school-by-school means or medians are provided.

In contrast to the present concern for "public information," consider the need for "public understanding." The data presently made available by school systems may promote the former, but they contribute little to the latter. The difference is a question of values, which we shall discuss in a moment.

Allocation of resources, modification of educational programs, and promotion of public understanding, then, are three ways in which test results can be used by education managers. We shall next consider the ways in which these ends can be achieved.

Paradigms for the Use of Institutional Test Results

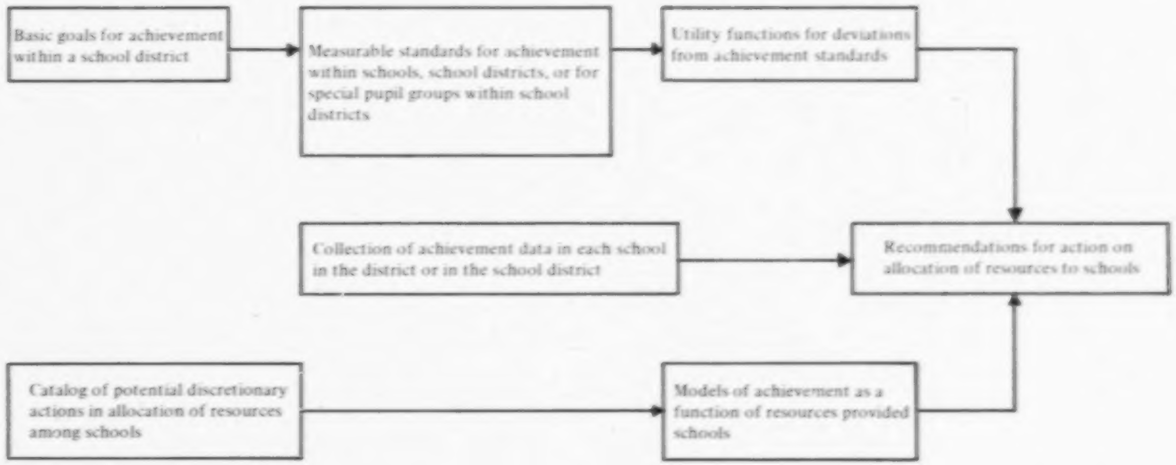
RESOURCE ALLOCATION

Discussing the use of test data in the allocation of resources among schools can be simplified by referring to Figure 1.

The use of achievement test results to guide resource allocation

Figure 1

A Paradigm for Using Test Results to Guide Allocation of Resources among Schools


Basic goals for achievement within a school district
Measurable standards for achievement within schools, school districts, or for special pupil groups within school districts
Utility functions for deviations from achievement standards
Collection of achievement data in each school in the district or in the school district
Recommendations for action on allocation of resources to schools
Catalog of potential discretionary actions in allocation of resources among schools
Models of achievement as a function of resources provided schools

requires the identification of fundamental goals for pupil achievement. Attainment of these goals is sought by allocating resources purposefully. Therefore, one scheme for resource allocation will be preferred to another if it results in a higher probability of realizing fundamental goals.

Goals for achievement can take many forms, depending upon the interests and values of the goal-setters. Some examples are: goals maximizing average pupil achievement in basic skills; minimizing the proportion of pupils achieving below some criterion value; or maximizing the proportion of pupils whose achievement exceeds some criterion value.

To be useful in decision making, goals must precisely define measurable standards through which reality and desired status can be compared. Bloom (1, p. 22) defined criteria for usable standards, which he termed "specifications":

> If education is to be open, public and examinable, the specifications for it must be explicit, and either the process of education or the outcomes of the process must be examinable in relation to such specifications.

Standards do not define utopian conditions, but conditions which are considered acceptable. They are necessary if one is to compare data indicating current status to conditions defining where one wants to be.

The kinds of educational goals so often tolerated do not lead to measurable standards. For example: "Each child should be allowed to develop to the fullest extent of his capacity." Such goals not only fail to define how full is full, but also do not permit the quantification of fullness.

To use tests to guide the allocation of resources (and for several other decision-making processes), utility functions for deviations from achievement standards must be defined. A statement often found in school test reports goes something like this: "The median reading achievement of sixth-graders in Middleville is 0.2 grade equivalent units below the national norm." Such statements are sometimes followed by exhortations to do better, implying dissatisfaction with test results. Utility functions quantify such dissatisfactions.

For example, consider only the matter of range of dissatisfaction. The citizens of a community might rightly be indifferent to an achievement average within 0.2 grade equivalent units of a national norm, mildly concerned if the difference between local achievement and the norm is reported as 0.3 to 0.5 units, and outraged if the mean drops more than 0.6 units below the norm. If such indifference, mild concern, and outrage could be quantified, one would have an index of the seriousness of educational problems.

To allocate resources wisely, education managers must consider all of the corrective actions available to them. The potential courses of action that education managers might take differ among school systems. In some systems, teacher education and experience can be considered an assignable resource, since teachers are centrally assigned to schools. In other systems, district officers allocate teacher positions among schools, which school principals then fill. In these cases, teacher experience is not an allocable resource.

Finally, education managers must know how resource allocation decisions will affect pupil achievement. If, in a particular school, knowledge of the humanities is far below standards and the utility attached to this deficit is large and negative, the decision maker must know which of his available resource allocation options will best remedy the situation. Should he assign more teachers, and thus reduce the size of humanities classes? Should he expend funds on in-service training for teachers already in the school? Should he purchase a new multimedia curriculum package? Rational decisions among such alternatives can only be made with knowledge of the probable results of each.

If all of the components of the resource allocation paradigm are available, reasonable policy can be formulated quite simply. A decision maker need only look at the potential actions available to him and choose those actions which provide the largest increments in utility with the highest probabilities. Or, alternatively, choose those actions which have the highest probabilities of alleviating the most serious problems.

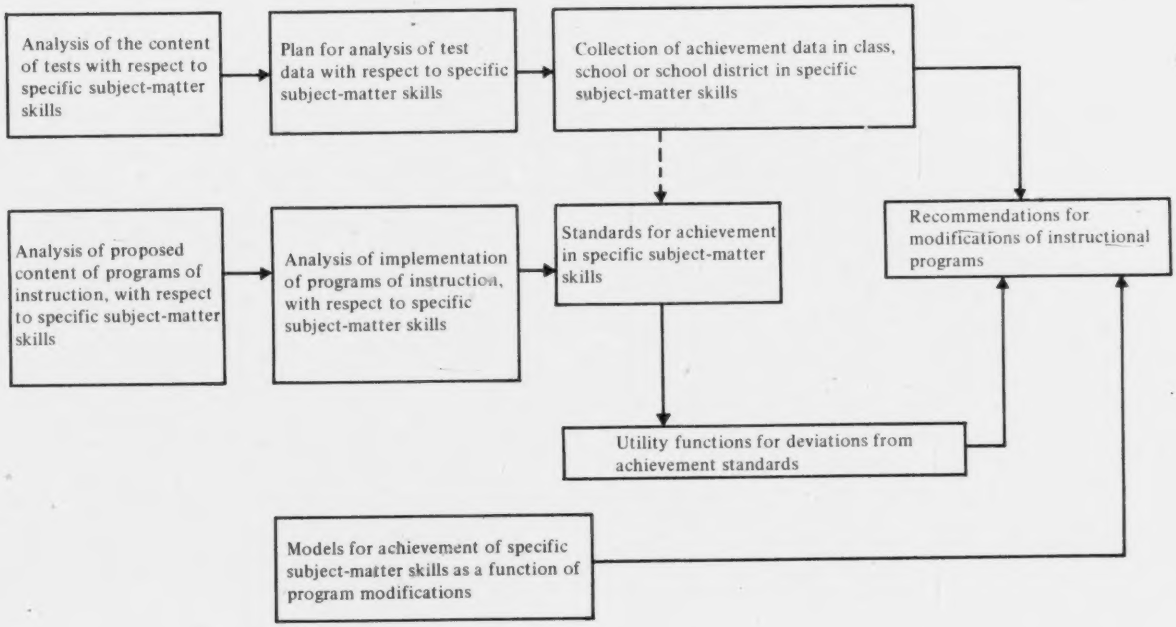
Some of the components of this paradigm are well within the state of the testing and management arts; others present problems requiring a complete redirection of our testing programs and our interpretations of test data. Paradigms for other applications of test results will be considered next, before discussing some implications for testing programs.

PROGRAM MODIFICATION

The paradigm for using achievement test results for decisions on program modification, shown in Figure 2, bears some similarity to that for decisions on allocation of resources. The program modifica-

Figure 2

A Paradigm for Using Test Results to Guide Modification of Instructional Programs


Analysis of the content of tests with respect to specific subject-matter skills
Plan for analysis of test data with respect to specific subject-matter skills
Collection of achievement data in class, school or school district in specific subject-matter skills
Analysis of proposed content of programs of instruction, with respect to specific subject-matter skills
Analysis of implementation of programs of instruction, with respect to specific subject-matter skills
Standards for achievement in specific subject-matter skills
Recommendations for modifications of instructional programs
Utility functions for deviations from achievement standards
Models for achievement of specific subject-matter skills as a function of program modifications

tion paradigm assumes the specification of standards for achievement in specific subject-matter skills. The attainment of these standards is assumed to be the objective of the instructional program. All components of the paradigm require analysis with respect to these subject-matter skills.

Commercially available achievement tests assess a multitude of skills under the same title. For example, at upper elementary grades, a reading subtest may measure word recognition skills, the ability to discern meaning from sentences, the ability to draw inferences from prose, and the ability to integrate information and arrive at a correct conclusion. Some of these skills undoubtedly relate more directly than others to the curriculum for which modification is being considered. The first step in using test data for program modification is the identification of those skills the program seeks to develop. Since instructional programs often differ from blueprint to implementation, analyses of the actual program—as well as the program blueprint—are required. When the specific objectives of the program have been identified, one must set standards for success against which achievement results can be interpreted. Analysis of the content of tests used in a school testing program should yield items that directly assess the skills the program seeks to develop and items that assess related, but secondary, objectives.

As in the paradigm for resource allocation, rational program modification requires the development of utility functions for deviations from achievement standards. Deviations that carry positive utility or small negative utility would probably not require modifications of programs. Deviations that carry large negative utility would imply the need for program modification, with specific decisions determined by analyzing models of achievement as a function of program change. Since very few educational inputs guarantee specific outputs, models of achievement as a function of program change would no doubt be probabilistic, for example, "If the length of training in word recognition is doubled, correct identification of 80 percent of a list of 400 fourth-grade words will be achieved with probability 0.9."

As with resource allocation decisions, program modification decisions are easily made when all of the components of the paradigm are available for consideration. The education manager attempts to treat situations that show a large negative utility. In treating those situations, he chooses program modifications that have the highest probability of success.

PUBLIC UNDERSTANDING

A paradigm for promoting public understanding of achievement in schools is shown in Figure 3. To inform the public of the status of achievement in the schools requires the collection, reduction, and analysis of test data, in addition to reporting. Statements such as "The median language achievement of fourth-graders in Middleville is at the forty-fifth percentile on national norms" inform the public of achievement status. However, such statements do not promote public *understanding*. Most educators would be quite cautious in interpreting this statement on median language achievement. Are Middleville fourth-graders doing reasonably well and parents need not be concerned, or are these students seriously deficient in language achievement? To understand the meaning of statements on achievement status, the public must be provided with, or helped to specify, a utility function. Again, standards for achievement consistent with broad educational goals must be clearly specified. Utility functions for deviations from standards are necessary to answer "how bad is bad" and "how good is good."

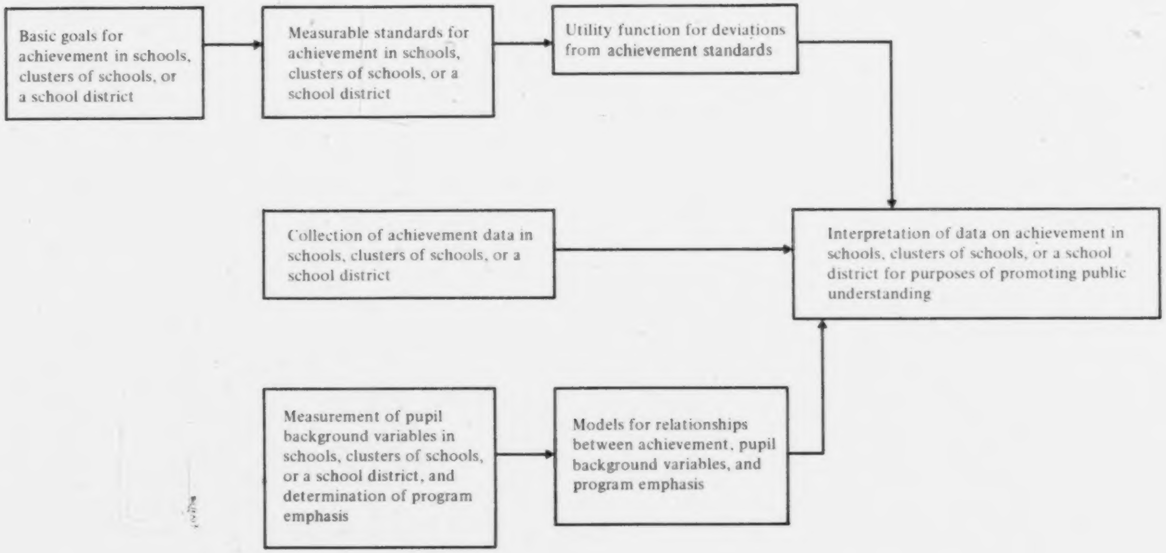
To promote public understanding of the meaning of test results, more than the scores themselves must be reported. Studies such as those of Burkhead (2) and the Office of Education's survey on *Equality of Educational Opportunity* (12) have shown significant relationships between pupil achievement and a host of pupil background variables. Affirming the generalities of these studies—that the economically poor are the academically poor, and that minority children achieve less well than majority children—is not sufficient. These may be realities, and the public should understand the extent to which they exist in local communities. But, more important, the public should be made to understand how the school system is *treating* such realities, what special programs are being implemented, and where.

In using the paradigm for promoting public understanding, we assume that the education manager will report to the public not only the status of pupil achievement, but the utility of that status relative to agreed-upon standards. Additionally, we assume that the public will be given an explanation of the probable causes of reported achievement and the school system's intended actions in response to the report.

These, then, are the paradigms, components, and their relationships which would permit institutional test results to be used effectively—for allocating resources among schools, for modifying educational pro-

Figure 3

A Paradigm for Using Test Results to Promote Public Understanding


Basic goals for achievement in schools, clusters of schools, or a school district
Measurable standards for achievement in schools, clusters of schools, or a school district
Utility function for deviations from achievement standards
Collection of achievement data in schools, clusters of schools, or a school district
Interpretation of data on achievement in schools, clusters of schools, or a school district for purposes of promoting public understanding
Measurement of pupil background variables in schools, clusters of schools, or a school district, and determination of program emphasis
Models for relationships between achievement, pupil background variables, and program emphasis

grams, and for promoting public understanding of the work of the schools. I would now like to discuss some of the components in greater detail and consider their relationships to the tests which are used and the ways in which data are interpreted.

Goals and Standards

Goals for education and standards for achievement are necessary components of all three paradigms. Dunkel (3) suggested that universal goals for education do not exist in our pluralistic society, and that school boards—the nomothetic proponents of goals—are intentionally vague in their formulations. In contrast to this view are the findings of Merwin and Womer (8), who noted a striking degree of agreement among school personnel, university professionals, and laymen on important goals for American education.

Goals for education and standards for achievement are implicit in present testing programs. By judging schools on the basis of commercially available tests, one sets as goals the development of those skills the tests seek to measure. Further, one sets as standards the median of scores achieved by the publisher's norms sample. How appropriate are these goals and standards? The answer lies in the structure and content of curriculums in the school systems where tests are used, and in the composition of the pupil population in those school systems. The acceptance of median national performance as a standard carries with it, first, the assumption that the test in question is as appropriate to the curriculum in a given school system as it is to the great diversity of curriculums encountered across the nation. Second, we must assume that the children in a given school system are, in their interests, abilities, and aptitudes, like those in the publisher's norm sample. I suspect that in many situations these assumptions are unwarranted.

Utility Functions

In addition to requiring well-defined standards for achievement, all these paradigms require that utilities be assigned to deviations from standards. Of those components needed to utilize institutional test results, the greatest deficiency probably lies in methods of deriving utilities for performance. I shall borrow an example from Ebel (4) to illustrate the problem. Suppose one were to construct a word meaning test in which words were systematically selected from a specified dictionary along with their meanings. Suppose words and meanings were listed alphabetically and students were instructed to match words with their proper meanings. Assume a standard of 60 words correct on a 100-item test. What utility should be attached to a median score of 45 words correct? Obviously, the utility should be negative; but how large should it be? Large enough to justify an expenditure for a remedial program? So small that a school district's curriculum designers can ignore the discrepancy?

Present testing programs and the typical interpretations of their results provide no utilities. Perhaps we are not yet used to interpreting test results for groups, where differences from implicit standards are not exceeded by the standard errors of scores. School district averages which differ by 0.2 grade equivalent units or 4 raw score points are, with high probability, statistically different. Are they substantively different? Most of us can't answer that question, hence we cannot attach utilities to deviations from standards.

Models of Relationships to Achievement

To allocate resources intelligently, educators must know the probable effects of their resource allocation decisions. That is, they must know the probable relationships between the availability of resources and desired educational outcomes. Similarly, to make the right decisions on the modification of educational programs, the probable effects of these decisions must be known. Again, knowledge of a relationship between actions and achievement is implied. Finally, public understanding of achievement test results requires information on the relationships between achievement test scores, other characteristics of pupils, and the structure and content of programs operating in the schools.

Knowledge of some of these relationships is scant, and decision makers are forced to operate with significant uncertainties. However, through the paradigms proposed here, the areas of uncertainty can be identified and perhaps researched. In making decisions, degrees of uncertainty can be treated as data and used to influence changes from the status quo.

Implications

We have identified more problems than solutions. It is clear that current testing programs do not provide education managers with the kind of information they feel is useful. It is also clear that there are many gaps to be filled in the paradigms suggested for using school test data.

I would suggest that content-standard tests, as proposed by Ebel, would be more useful for identifying standards and utility functions than the norm-referenced tests now commonly used in our schools. In this approach, a domain of test items can be explicitly linked to goals and curriculums. It is, for example, far easier to identify a standard for such specifics as "word knowledge" or "ability to understand the meaning of sentences in prose" than to derive a standard for overgeneralized "reading" or "language skills."

Perhaps norm-referenced tests serve well for individual guidance and the array of decisions teachers and counselors must make in assisting individual children. However, there is no reason why a single testing program must be used to serve both the needs of individual decision makers and institutional decision makers. Separate but complementary testing programs might best meet these separate but complementary needs. For institutional decision making it is surely not necessary to test every child. Nor is it necessary that every child complete the same test items. Research conducted this past year leads me to suggest that many institutional decisions can be based upon test results for as few as five percent of the children in a school system, provided these children are sampled correctly. The resulting economies will permit a much broader range of testing than is now possible within constrained budgets, and will permit use of testing methods that would not be feasible if tests were administered to all children.

Whatever testing models we employ, it is clear that we must make more explicit our reasons for testing and the intended use of results—for our own benefit as well as that of the education consumer. It is also clear that we have a large task ahead if we are to properly *utilize* testing data.

REFERENCES

1. Bloom, Benjamin. Some theoretical issues relating to educational evaluation. 68th Yearbook of the National Society for the Study of Education, Part II. Chicago, 1969, 26–50.

2. Burkhead, Jesse. *Input and output in large city schools.* Syracuse, N. Y.: Syracuse University Press, 1964.

3. Dunkel, H. B. Value decisions and the public schools. *The School Review*, Summer 1962, *70*, 2, 163–170.

4. Ebel, Robert L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 1, 15.

5. Friedenberg, Edgar. Social consequences of educational measurement. *Proceedings of the 1969 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1969.

6. Guthrie, James, and others. *Schools and inequality.* (in press) (1970).

7. Haggerty, M. E. Specific uses of measurement in the solution of school problems. *The measurement of educational products:* 17th Yearbook of the National Society for the Study of Education, Part II. Bloomington, 1918. 25.

8. Merwin, J. C. and Womer, F. B. Evaluation in assessing the progress of education to provide bases of public understanding and public policy. 68th Yearbook of the National Society for the Study of Education, Part II. Chicago, 1969, 305–334.

9. Schrag, Peter. End of the impossible dream. *Saturday Review*, September 19, 1970. 68 ff.

10. Sexton, Patricia. *Education and income.* New York: Viking Press, 1961.

11. Tyler, Ralph. The functions of measurement in improving instruction. In Lindquist, E. F., (Ed.) *Educational measurement.* Washington, D.C.: American Council on Education, 1951. 47–67.

12. Coleman, James S., *et al. Equality of educational opportunity.* Washington, D.C.: Department of Health, Education and Welfare, Office of Education, 1966.

National Assessment

ROBERT E. STAKE
University of Illinois

The way I see it, National Assessment is Ralph Tyler's baby. Some folks call it Frank Keppel's baby. Some folks call it Wendell Pierce's baby. Some people think it's Rosemary's baby. But I still see it as Ralph Tyler's baby.

Several years ago, when Ralph Tyler spoke to educators and government officials about the plans for National Assessment, he talked about "Indicators of Educational Progress." He talked about a "Gross Educational Product," somewhat equivalent to the Gross National Product. He said that indicators would help the educational leaders of the nation set policy and assess the progress of our teaching and learning.

Later, when Ralph Tyler talked to the many subcommittees of the Exploratory Committee for Assessing the Progress of Education, he charged them with responsibility for stating the nation's educational goals. They got busy and wrote objectives in 10 subject-matter areas.

And when Ralph Tyler talked to Jack Merwin and Frank Womer about implementing National Assessment, he charged them with developing an information system. The principal elements of this system would be performance exercises. The exercises would reflect those previously-stated national objectives. Content validity would be stressed. Each science exercise, for example, would be meaningful alone, not needing to be grouped together with other exercises to make a science score. And they did what Ralph told them to do—they started the information flowing.

First Reports to the Public

The first results of National Assessment were announced at a meeting called by the Education Commission of the States (ECS), National

Assessment's sponsor for a year now. A national sample of about 100,000 children and adults had responded during the previous year to a total of about 460 exercises in science, citizenship, and writing. Each person—according to the matrix sampling plan—was tested for less than an hour. At the ECS meeting last July in Denver, information on science and citizenship exercises was released.

Tables 1 and 2 show the results of two science exercises. Table 1, of exercise number 222 results, tells us—crudely—how many children

Table 1

EXERCISE 222

20% difference in favor of age 17

In terms of the theory of natural selection, what is the explanation of why giraffes have come to have such long necks?

Age 13	*Age 17*	
8%	12%	Stretching to get food in high trees has made their necks longer.
2	1	There is something inside of giraffes which keeps making longer necks.
12	6	Giraffe food contained vitamins which caused the vertebrae to lengthen.
28	13	Giraffe necks have gotten longer as time has gone on, but nobody has any idea why this is.
38	58	Giraffes born with the longest necks have been able to stay alive when food was scarce and have passed this trait on to their offspring.
12	10	I don't know.
0	0	No response.
100%	100%	

At age 13, "the" was omitted from the third alternative.

(The above exercise was taken from National Assessment of Educational Progress Report No. 1, "Science—National Results," July 1970.)

can use Darwin's theory of natural selection to explain why giraffes have long necks. To respond correctly, the child needs to know the main principle of this theory and must apply it to the case of the giraffe. (Or he needs to know that the longest alternative is the best in a multiple-choice test.) Thirty-eight percent of the 13-year-olds chose the right answer. National Assessment people did not interpret this finding—nor was such an explanation part of their game plan. They reported what the national percentages were, and the implication of these percentages is to be left to national and local educational leaders.

The exercise information shown in Table 2 is from an individually administered exercise using apparatus. It shows the percentage of correct responses for region, size of community, type of community, sex, race, and parents' education level.

These two exercises are not representative of the total pool of exercises—perhaps no two items could be. Most of the science exercises are multiple-choice items requiring a knowledge of factual information. The citizenship items that have been released are, for the most part, open-ended questions requiring a self-estimate of typical behavior. For example, 13-year-olds were asked if they would step forward to protest a particular example of racial discrimination in a public park. Results for the national sample on 35 citizenship exercises have been released to date; more of the same, and results of exercises on writing, will be published in the near future.

Criticisms of National Assessment

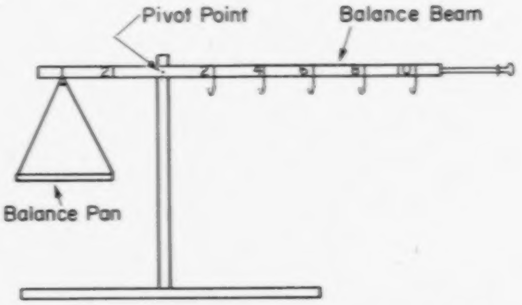
What has been the reaction to this information? Most public officials and professional people seem to be saying, "Yes, this National Assessment is something Education ought to be doing, but . . ." Two of these "buts" are:

> ". . . but why did they weigh factual knowledge so high and advanced understanding and learning skills so low, in this collection of exercises?"
>
> and:
>
> ". . . but why aren't the National Assessment people telling us what these percentages mean? Will their information tell us whether the education climate is stormy or sunny?"

I should not imply there has been a recent wave of criticism. Articles

Table 2

EXERCISE 341



"The apparatus before you is the same as that shown in the picture. This balance is balanced when the balance beam is level as shown above. The number by each mark on the beam tells the number of inches that mark is from the pivot point."

"Place one weight in the balance pan. How many inches from the pivot point is the hook on which you must hang one weight to get balance?"

	*Number in Group**	*Percent Correct*	*Difference: Group vs. All 17s*	*Approximate Standard Error of Difference*	*Difference Divided by its Standard Error*	*Statement About the Difference*
All 17s	593	75.2%				
Region						
Northeast	159	84.0%	8.8%	3.02%	2.91	perhaps 9% higher
Southeast	124	64.7	−10.5	5.10	−2.06	perhaps 11% lower
Central	144	70.1	−5.1	3.20	−1.59	may be lower
West	166	81.0	5.8	3.01	1.93	may be higher

(*Continued on next page*)

	Number in Group*	Percent Correct	*Difference: Group vs. All 17s*	*Approximate Standard Error of Difference*	*Difference Divided by its Standard Error*	*Statement About the Difference*
SOC						
Big City	148	70.5%	−4.7%	3.30%	−1.42	—
Fringe	139	82.3	7.1	3.05	2.33	perhaps 7% higher
Medium City	153	72.5	−2.7	3.59	−.75	—
Smaller Places	153	72.7	−2.5	4.15	−.60	—
TOC**						
Inner City	60	73.9%	−1.3%	5.78%	−.22	—
Suburb	53	76.9	1.7	5.64	.30	—
Rural	52	73.5	−1.7	8.55	−.20	—
Remainder	405	75.3	0.1	1.01	.10	—
Sex						
Male	299	81.1%	5.9%	2.12%	2.78	perhaps 6% higher
Female	294	68.8	−6.4	2.36	−2.71	perhaps 6% lower
Color						
Black	84	50.0%	−25.2%	5.80%	−4.34	roughly 25% lower
Non-Black	509	77.9	2.7	0.65	4.15	roughly 3% higher
Parents' Education						
8th or below	56	67.9%	−7.3%	8.24%	−.89	—
Some High School	102	64.5	−10.7	5.30	−2.02	perhaps 11% lower
HS Graduate	173	77.3	2.1	3.12	.67	—
Post High School	250	79.3	4.1	2.18	1.88	may be higher

*The package in which this exercise appeared was individually administered. Individually administered packages were given to a smaller sample of respondents than were group administered packages; hence, the smaller sample sizes shown on this page.

**All in-school 17s = 75.2%.

(From National Assessment of Educational Progress Report No. 1, "Science—National Results," July 1970.)

in *Time*, *Newsweek*, *U. S. News and World Report*, numerous newspaper articles, and statements by educational leaders have been indulgent, if not enthusiastic. "Criticism of National Assessment has disappeared," claimed Martin Katzman and Ronald Rosen of Harvard, who then proceeded to fill the void. The consensus of criticism—what there is—is aimed at the blandness of the objectives and the emphasis on factual knowledge. In this paper I, too, will complain about the objectives, but I will belittle those who complain about the emphasis on factual knowledge.

These two criticisms alert us to the breach between Ralph Tyler's three conversations—about national objectives, about exercises, and about indicators of progress. How do they tie together? How do you satisfy yourself that the chosen objectives are the right objectives? And how do you satisfy yourself that the chosen exercises are valid indicators of the objectives?

Unified Objectives in a Pluralistic Society

Table 3 illustrates a committee procedure for selecting National Assessment objectives that is almost guaranteed to produce bland objectives.

Dick Jaeger mentioned the article "The End of the Impossible Dream," in which Peter Schrag said,

> Any single, universal institution—and especially one as sensitive as the public school—is the product of a social quotient verdict. It elevates the lowest common denominator of desires, pressures, and demands into the highest public virtue. It cannot afford to offend any sizable community group, be it the American Legion, the B'nai B'rith, or the NAACP.

This is exquisitely true of National Assessment. The decision to filter all objectives through a committee of subject-matter experts, a committee of educators, and a committee of citizens yields a product that even an ulcer-ridden public can find inoffensive. To mollify the public and the profession may be good politics, Katzman and Rosen remind us, but it does not discharge our professional responsibility to attend to the concerns of minority groups, curriculum innovators, social planners, and silent-majority folks. The present objectives are

Table 3

Groups convened by the National Assessment Project for the purpose of passing judgment on the objectives for which exercises will be developed:*

1. *Subject-matter specialists.* Specialists in the subject area must consider the objectives authentic from the viewpoint of the discipline. Scientists must agree the science objectives are authentic; mathematicians must agree upon the authenticity of the mathematics objectives, etc.
2. *Educators.* School people must recognize them as desirable goals for education and ones which schools are actively striving to achieve.
3. *Citizens.* Parents and others interested in education must agree the objectives are important for youth and young adults to know, feel, or understand.

The current National Assessment objectives in the area of Science are:**

1. Understand the investigative nature of science
2. Possess the abilities and skills to engage in the process of science
3. Know the fundamental facts and principles of science
4. Have attitudes about and appreciation for scientists, science, and the consequences of science that stem from adequate understanding

The current National Assessment objectives in the area of Citizenship are as follows:**

1. Show concern for the well-being of others
2. Support rights and freedoms of all individuals
3. Recognize the value of just law
4. Know the main structure and functions of our governments
5. Participate in effective civic action
6. Understand problems of international relations
7. Approach civic decisions rationally
8. Take responsibility for own development
9. Help and respect their own families

---

*Taken from page 12 of *The National Assessment Approach to Exercise Development* by Carmen J. Finley and Frances S. Berdie. Ann Arbor: National Assessment of Educational Progress, 1970.

**Taken from National Assessment of Educational Progress Reports No. 1 and 2, *Science: National Results* and *Citizenship: National Results—Partial.* Denver: Education Commission of the States, both July 1970.

wanting, and the present procedure for selecting objectives is impotent.

Katzman and Rosen are pessimistic; they do not expect National Assessment objectives to be improved. I am optimistic. I do. I think that the National Assessment people can be persuaded to ditch the one-track, universally-acceptable-objectives-only model and set up a procedure to solicit and use innovative, and parochial, and anachronistic exercises that do measure somebody's goals. Something for everybody? Why not? The advantages are clear. Honor the pluralism of our people; increase the face validity of the collection; and emphasize that it is up to individual teachers, school board members, citizens, and national officials to decide—not up to the National Assessment staff or its committees to decide—what objectives and exercises to pay attention to in making educational policy.

There is a great deal of merit, as Ralph Tyler has told us in his writings over 40 years, in orienting a curriculum or a testing program around the purposes of education. But there is also a great danger that the purposes we measurement people identify will be distortions of what our colleagues are saying and irrelevant to many of our constituents.

National Assessment has an obligation to encompass more than the popular, the inoffensive, and the easily and reliably measured. How can it do better? Not through better committees, I believe, nor by lifting its sights to the higher aims of education. The quality control of objectives can be accomplished by a good empirical-data feedback system. Try them out in the field. Which satisfy? Which objectives and which exercises have a constituency? What information has an audience? These are the questions you ask to find the right objectives and exercises for National Assessment.

Indicators of Progress

The second present criticism was aimed at needed interpretations. Are exercises taken alone to be the primary Indicators of Progress? If not, how do we assemble a good barometer? What is a Gross Educational Product, anyway?

[At this point, without further announcement, a film was projected onto a large screen behind Dr. Stake and the following script was acted out. The scene was the weather desk of a television studio. The "weatherman" speaks.]

"Good evening: Your Education Report is brought to you by Peerless Performance, Incorporated, the friend of your school, the friend of your child.

"Reading achievement continued to dominate the national picture this past month. Along the eastern seaboard, achievement rose to the high 80s for the third month in a row.

"A low-pressure system—de-emphasis on reading—which last month centered on Texarkana, Texas, has been moving up the Ohio Valley. This system may eventually bring the eastern readings back to normal.

"Experts continued to watch the drop in adult reading scores in the Miami, Florida, area. The scores on such exercises as the Gates Group showed annual losses up to 15 percent. The Gates Group exercises feature newspaper articles such as this one that the learner must read and interpret.

"Here in our own community, the all-group reading mean remained at 82. The datatronic curriculum continues to lower the priority on mathematics, leaving us with only 148 learning-day equivalents for the year. The social-group-pressure indicator stands at 13 in the public schools, 15 in the private. During the past 10 school days, only on one day did the distraction quotient rise above 20 percent.

"On the big board we find the Gross Educational Product at 748. The trend continues to show a rise of about 10 points per year."

Though exaggerated here for effect, this glimpse into the future is really not so farfetched. People are acquainted with indicators. It is estimated that almost half the people of the nation listen to at least one weather report, via radio or television, every day. They complain about the weather; they complain about the accuracy of weather forecasts; but they seldom complain about the dazzling array of variables on which the meteorologists have chosen to report.

I am not going to contend that the weather report enables its information consumers to make rational decisions. Maybe it does, maybe it does not. I do want to make the point that people accept the indicators of the weather and incorporate them into their communications. Educational indicators should shoot for such a goal. If their indicators should also become a staple of conversation, I believe they will influence good planning.

Moreover, weather reports and other popular presentations of statistics can suggest several keys to the effective presentation of educational information. In the educational weather report we just viewed, the information was presented with accentuated reference to space and time. Geographically-based information and time-based comparisons are easy to understand. Time and space are powerful dimensions for generalization. What *has* happened may continue to happen. What is happening *here* may happen over *there*. Thus, it seems to me that an educational-assessment system is likely to rely heavily on regional and temporal indicators to convey the picture of educational progress.

National Assessment has been designed to give us a graphic plot of progress through time in different geographical areas. Developing the time dimension will try our patience because, according to the present National Assessment schedule, it will be at least six years before three points can be plotted as the beginnings of a trend line. That schedule should be altered to permit at least a few indicators to be plotted annually, or even oftener.

The only space comparisons currently promised by National Assessment are regional, with only four regions to the nation. However, there is an increasing demand for state-by-state and district-by-district comparisons. Legislators and citizens' committees are probing for criterion information by which good schools and bad schools might be identified. "Ouch," say many school men. They feel that National Assessment exercises won't be the right criteria, or certainly not *all* the right criteria, for evaluating the quality and productivity of the schools. They have had tests around for 40 years and have little reason to believe that tests or exercises or assessment indicators will show what the teachers are teaching. They don't want evaluation, at least not if it is based upon criteria other than their own. They don't want any school-by-school comparisons that have been proposed so far.

The National Assessment staff also says "Ouch" to this demand—but they have a different reason for not tooling up to provide within-state comparisons. They want a grace period, a chance to work unhampered, a chance to demonstrate that they can provide useful information. If called on to assist states and local communities in assessment, and if obligated to defend their initial choices of exercises before every challenger, before every disgruntled teacher, before every below-the-norm school district, and before every visionary social critic, they will not be able to give National Assessment a fair try. So even though ECS has stated that it will assist State Assessment activities, the National Assessment staff hopes not to be involved.

Each of the states initiating state assessments—for example, Colorado, Florida, Massachusetts, and Nebraska—could render a great service by defining its own indicators of progress. It would be nice to have a Florida indicator of science achievement, and a Michigan indicator, to remind people that no one indicator is the whole truth. Conceivably, after a while people would become aware of the sensitivity of certain indicators to things that are important in their observations of education; they would rely on some indicators, and others would fade from their scene. Some indicators would fade from all scenes, for the same reason that giraffes have long necks.

I am quite serious in thinking of indicators in Darwinian terms. More than ever before, communication is a jungle. Only the fittest messages will survive. The indicators that people pay attention to, that become part of them and useful to them, will survive. And this survival may be unrelated to the quality of the information they contain.

The Gross Educational Product

As far as I know, nobody has any good idea of what the ingredients of a Gross Educational Product should be. The GEP obviously should be a composite of information about many dimensions of education. Should it be limited to the basic knowledges and cognitive skills? Should it include something from the affective and psychomotor domains? Should it include the educational productivity of adults? Should it include what is primarily learned in nonschool settings, such as in the locker room, in the barracks, in the shopping center, and in the family car?

The National Assessment staff and their subcontractors have not yet drafted even a first sketch of an answer to these questions. They should. And others of us in educational measurement should. It is an important technical area within our jurisdiction, though not within our present competence. But I would suggest that answers to the question of ingredients are not as critical as most people think. I would argue that the value of a social indicator may not be closely related to the importance of its particular ingredients, that an indicator based on growth in factual knowledge may be more valuable for educational planning than an indicator based on higher mental processes.

There will be a denotative meaning and a connotative meaning to every indicator. Its formal definition—its composition—may be one thing; but its meaning in informal discourse will be another. The present Gross National Product, I am told, is ridiculed by some economists because it is based, they say, on a poor choice of ingredients. But as a citizen and consumer, I could not care less. The GNP is a useful indicator to me and to many economists and political leaders. An indicator will be useful to me if it correlates with things in my experience.

If I can persuade you to remember but one thing at this time, let it be this: that a continuing assessment of educational progress creates its own meaning of progress. We are not clairvoyant. We cannot forecast tomorrow's meaning, the clinical meaning of Gross Educational Product.

What we should do now is worry about National Assessment ranging far enough. Will the more parochial and complex and exotic bits of information become available as possible ingredients? We should not become bogged down in planning the ideal indicators of progress but should try many of them, knowing that the fit will survive.

And so in conclusion, I would reiterate—

[At this point a recorded voice interrupted Dr. Stake]

Voice: Now, wait a minute.

Stake: Did someone have a question?

[Projected on the screen behind him, the filmed figure of Stake himself comes through a door from a lighted hall into a darkened room, then turns the lights on.]

Stake on screen: Yes, I have a question.

Stake at podium: What is your question?

Screen: Aren't you ignoring the real issue? This talk about when indicators become meaningful, and whether or not to emphasize the Establishment's objectives or to upgrade the objectives . . . it doesn't answer the question "Is any assessment any good?" How about it? "Is National Assessment part of the promise or part of the peril?"

Podium: How can you answer a question like that?

Screen: Well, you can try. An awful lot of people think that education in the U. S. is in a hell of a fix. You didn't even mention that National Assessment might be a poison apple, beautiful to behold with its item sampling and content validity, but the kiss of death to creative teaching. Aren't you going to consider the possibility that National Assessment might aggravate our problems and help blind us to the important responsibilities of our schools?

Podium: What do you think the specific peril might be?

Screen: If I were Marshall McLuhan, I would say something like: Testing is the medium is the message. It's not what we learn from testing that counts but what we tell by testing. We tell what *we* value, what we think is important. Peter Caws, who recently referred to National Assessment in the *New Republic*, doesn't seem to get any other message from it than: To educators an educated man is one who recognizes as true one sentence out of four. Isn't National Assessment more peril than promise because it encourages people to think that education is much less than it really is?

[The film figure fades from the screen and Dr. Stake resumes speaking.]

Yes, we should shudder at least once again about the peril in National Assessment. Each new step is perilous, toward the moon or toward the next generation of technologies. There is potential peril in every measurement, in every testing program, in every effort to get a better understanding, in every effort to communicate. Each measure has its error, every social venture has its side effects. Any one error may tip the balance from a good to a poor choice, from a wise to an unwise national investment. National Assessment is an effort to simplify and bring within the reach of our understanding the robustness of education in this nation. It cannot help but be an oversimplification.

But we know that every index number, every graph, every word of prose is an oversimplification. We have no choice but to create simple things to stand for complex ones. Our curiosity, our desire to command our destiny, demands it. We are human beings. We will not be persuaded that it is wrong to define, to symbolize, to model, to measure.

National Assessment today is at the beginning of a massive, expensive field trial; a reasonable evaluation of its utility cannot be made before 1975. We can take some comfort in 1970 in the fact that its staff is honest, competent, and productive.

Thank you, ladies and gentlemen and, in all respect and sincerity, thank you, Ralph.

Discussion

JAMES N. JACOBS
Cincinnati (Ohio) Public Schools

Frank Womer and I have divided our chores here—maybe, better yet, pleasures—in that I'm going to respond to Dick Jaeger's paper and Frank will be responding to Bob Stake's.

First of all, I thought that the topic to which Dick addressed his paper was extremely important to educators. It's obvious that the tight budgets that face most educational institutions mean that we have to scrutinize more closely the value of programs than we have ever done before. So I would agree with Dick's major thesis that institutional decisions have to rely more and more on test results, and, similarly, test results have to be used for more than just guidance purposes.

But why aren't they?

I'd like to advance a thought that is akin to Parkinson's law: The more weighty the decision to be made, the less reliance, or attention, to information.

If you agree with this notion, I would suggest that the reason—or one of the major reasons—is that tough decisions require a synthesis of many kinds of data. In Dick's paper he made the assumption that his system—his model system—was to be free of social and political factors. I would suggest that this is a very tenuous assumption.

As a matter of fact, we have to be cognizant of many kinds of information, not just test information. Each decision maker must figuratively write a mental equation with numerous information variables and their appropriate beta weights to yield the decision. Faced with such a task, I suspect that many would leave it to George, or ignore relevant information and make a seat-of-the-pants decision, based on what is sometimes called "experience."

Decisions about children are simpler than decisions about institutions. At least they are thought to be. Maybe this is why testing program information is not used as much in the category of institutional decisions by educational managers.

Dick put nonpupil decisions into the category of institutional decisions. Perhaps this concept ought to be brought into relief by suggesting the educational referents on which decisions are usually made, and further by suggesting the major users, or decision makers, of data describing these referents. I would suggest four target groups.

The children themselves are the first reference group. Tests are used for diagnosis, for placement, and so on, and the major users are the teachers, the pupils, pupil personnel specialists, and parents; and these people obviously are the decision makers.

Now, at this level, as Dick has suggested, tests find their greatest use—and I would add also, misuse.

The second reference group I would suggest is the teacher-class unit. Incidentally, the teacher and the class may represent two distinct referents, but my preference is to think in terms of collections of pupils, that is, the class.

Test results may guide class instruction, help identify needed resources, and so on. The major users of such data are teachers, supervisory personnel, and perhaps to some extent school administrators.

The third referent I would suggest is the school unit level. Test results may be used to make decisions on types of programs needed, school organization, whether or not remedial or enrichment classes are to be set up, and so on. Such decisions are usually made by the principal, but may also be shared by supervisory people or district directors and perhaps parent groups. Note that at this level we would speak of school averages of test results, just as at the class level we would be talking in terms of class averages. The uses described by Mr. Jaeger—that is, resource allocation, program modification, and public understanding—are most appropriate at this as well as the next level in this schema.

I would agree with Dick's thesis that current use at this level is very low.

The fourth referent is the school system. Decisions again can be made on resource allocation and program modification, and the results must be shared with the public at large. Superintendents, boards of education, and central office personnel are other major decision makers institutionally, but the public at large shares the information and at least sanctions, if it does not help make, the decisions.

Now, this schema represents successive aggregations of data, including test data, each with its special uses, users, and reference groups. The schema may be appropriate for all kinds of information, not just test results. Incidentally, in Cincinnati we are now developing a school unit level information system under a Title III project.

Dick's argument that the utility of test data rests on the specification of goals and setting of standards is very powerful. In my opinion, these tasks are among the toughest we face. I do not agree that there is public consensus on goals or standards for education. Agreement is found only at an abstract, highly conceptual level. The more we define our mission, the better the profession can accomplish its job—but at the same time the more the potential for resistance, both within the profession and without.

Our challenge is to construct banks of possible educational outcomes, permitting selection from among these outcomes to suit a given population. Many people must get into this act, not the least of whom is the individual student. The implication to testing, to my mind, is that norms reference measurements will be supplemented by criterion reference collections of test items tailored to the values, needs, and expectations of specific groups.

As to the need to set standards, I believe we must respond to the public demand for accountability on the one hand; yet on the other hand we must recognize that the human appetite for more of everything is insatiable. What is good or minimal has always baffled mankind, and probably always will. When standards are set, when they are valued by some group, and when they become measurable, they are ready for change. If this is not recognized by measurement people, we'll end up hopeless neurotics.

Now, just three more thoughts. First, the concept of utility function that was described by Dick as a deviation from status and standard has high heuristic value and should pose an enormous challenge to research and evaluation people. Eventually, the problem of measuring and weighting human values will have to be addressed.

I'd also point out that perhaps "problem index" might be a more descriptive term than "utility function."

Dick seems to have addressed himself to the basic elements of a much-talked-about but little-done-about program-planning budgeting system, except for one detail, and that is costs. To quote Dick: "Choose those actions which have the highest probabilities of alleviating the most serious problems." And to this I would add: "at the lowest cost."

Discussion

The need for public understanding, not just public information, of test results cannot be overemphasized. I doubt that the typical research or measurement people in public schools could do this job even if they had the time. We need a class of liaison people, between the community and the school, who are technically competent and who have extraordinary communication skills.

Among other things, they could serve the role of ombudsmen, thereby helping to bridge the credibility gap that has grown wider between the schools and the public.

My parting thought: Matching decisions with relevant data has a long way to go. We've got some nice data for which we don't have any decisions to make. We've also got some decisions to make for which we have no data.

Discussion

Frank B. Womer
National Assessment of Educational Progress and
The University of Michigan

When John Hayman* was introducing Bob Stake, he omitted one bit of information that might shed some light on the very fine stage performance Bob gave in presenting his paper. One of Bob's grandfathers was known professionally as Pawnee Joe and was known in the Denver area as the best Indian dancer in Dick's Wild West Show. So it's interesting to contemplate the fact that we have been privileged to have a performance by Pawnee Joe's grandson, better known as Illini Bob.

In chatting with Bob, I indicated some concern that he hadn't really given me very much to disagree with in his remarks. I do have, however, a couple of points that I would like to make.

This session is, I would hope, the beginning of a new dialogue on National Assessment, a dialogue concerned with looking at the National Assessment model and suggesting improvements. In the fall of 1970, National Assessment is fairly well established, yet we are not completely out of the woods. We have completed one full year of data collection. We are in the midst of our first year of reporting. We've started the second year of data collection. Our school district cooperation is 95 percent.

Some three to five years ago the major criticisms of the project came primarily from school administrators and were concerned with such things as the potential of a national testing program, curriculum domination, federal control, et cetera. Considering the fact that the criticisms that we had then have now resulted, three to five years later, in 95 percent school cooperation, I'm hopeful that the criticisms we're getting this year may in another three to five years result in almost complete acceptance by the educational community of what we are doing in the assessment project.

---

*John Hayman of the Great Cities Research Council, Chairman of Session A, "Educational Applications," of the Concurrent Sessions of the Invitational Conference.

Bob's major concerns, it seems to me, are twofold: first of all, bland objectives; and secondly, the need for educational indices. He does make some other points that I will react to, but these are the major points.

First of all, I think it is obvious from all of our publications that National Assessment's objectives are consensus objectives. Therefore, they do not include objectives held only by subgroups of the population. But in similar fashion, the Gross National Product itself is built only on selected inputs. Thus, I'm not sure that consensus objectives necessarily result in bland objectives, and I'm not really sure whether Bob is as concerned about the objectives as he is about the exercises. It seems to me that he suggests rather specifically more pluralism in the exercises. This actually can be accomplished without much change, if any, in the objectives themselves. In fact, we are attempting to move in that direction, although I must admit we have not done as much as we could do in terms of greater diversity in the exercises themselves.

But even under the present objectives, we can make progress in that direction. It's a direction in which we should be moving, and I certainly agree that our materials are not as diverse as they eventually should be.

Secondly, Bob asked for multiple indices of "gross educational product," developed both by National Assessment and by non-National Assessment personnel and groups. I couldn't agree more. I think that we within the staff should make an attempt to do this. I'm hopeful that Bob will make a similar attempt. And I'm hopeful that many conferees will attempt to develop indices based upon our results.

I don't think that this is a job that should be left entirely up to any single staff, not even the National Assessment staff. At the moment we are still struggling with the very first reporting. We have more coming up. We haven't even completed one round of what we would consider our basic reports. Our main effects and our interactions haven't even been computed yet. But certainly the time will come when we must begin to look at potential indices.

Incidentally, we use a somewhat different terminology within the staff. Bob has used GEP, gross educational product. A couple of months ago I prepared a memo for one of our advisory committees using GNK, for gross national knowledge—a memo related essentially to the concern about potential indices. John Tukey, chairman of that committee, suggested that we change that acronym slightly and use GRO from gross, N from national, and K from knowledge. Thus, within the project we are referring to this whole area as GRONKing. As we consider development of indices, the staff will attempt to GRONK, and we hope others will also.

Bob made a couple of other points that I might mention just in passing. There was a criticism of the factual nature of the exercises reported in July. This is true of the science area. But lest you consider that all of our materials are heavily weighted factually, I hasten to add that other subject areas are not as heavily weighted as science with factual exercises.

Bob commented that the staff feels a need to do its own thing, and I couldn't agree more. He stated it in much better fashion than I could. We do feel the need for time to attempt to follow through on the basic initial objectives of the project without worrying too much about additional tasks at this point.

In general, then, it seems to me Bob has had several suggestions for improvement and/or expansion of National Assessment—more diverse objectives and development of educational indices. Other commentators this year have asked for expansion to state assessment, for more complete studies of the achievement of various ethnic groups, for additions of new areas, and so forth. National Assessment is not yet—and, hopefully, never will be—a complete model. Changes can and should take place in the project. But in my opinion National Assessment should not be expanded or changed to handle every idea that is produced, even very, very good ideas. However, very good ideas should be explored carefully with the possibility that National Assessment might accommodate some of them, and that others should be handled through independent projects.

My hope is that there will be considerable spin-off from National Assessment to other investigations. I'm fearful of National Assessment being forced, because of political pressures, to add pet projects belonging to important people, a situation which might dilute its major thrust of developing into a national project for gathering information about educational outcomes. Such pressures already exist.

If the educational community, and specifically the educational research community, feels that it has a stake in National Assessment—

stake, that is, with a small "s"—it must expand the dialogue about what National Assessment is and should be. We must make our desires known if they are to be heard among the many that are being pressed upon National Assessment.

Session B

Technical Issues

Bayesian Considerations in Educational Information Systems

Melvin R. Novick
American College Testing Program and
University of Iowa

For many years students' scores on academic aptitude tests have provided selective colleges and universities with one important piece of information relevant to their decision of whether or not to select a particular applicant. Such tests have had the desirable effect of making admissions decisions for these institutions more dependent on academic promise and less dependent on status and influence. The result has been a broadening of the base of educational opportunity in this country. I am confident that these tests will continue, for some time, to serve this function.

Our educational system now, however, is in the process of redefining its constituency at the post-secondary level (16) to include essentially all students who can effectively benefit from any additional education (17, 9). This trend is best seen in the recent and projected growth in the number of students attending community colleges. One result of this trend is the growing number of students in nonselective colleges. Decisions of consequence for such students center largely in the choice of program of study.

Concomitant with this growth has been a broadening of the range of available educational opportunities. If this broadening continues, and if there is an increase in the diversity of training methods to accommodate students with different ability profiles, we shall approach a *meaningful* national policy of open admissions. This does not suggest that any one institution will need to encompass any greater range of programs or any greater number of students than it can effectively handle. It means only that the educational system as a whole will serve a much wider constituency.

In this situation it will be both possible and desirable to maximize the informed participation of each student in the decisions that affect his educational career (17). Indeed, to a very great extent the student, not the college, will be the primary decision maker. It will be the student who requires information about himself, the colleges, and the particular programs that may be relevant to his goals. In this context, educational testing becomes just one component of a decision-oriented information transmittal system having a guidance rather than a selection orientation.

Since 1964, the American College Testing Program (ACT) has provided a guidance-oriented information system, now used annually by approximately one million college applicants to both two- and four-year colleges and universities. This program provides the student with test scores and a variety of other information about himself. It also provides him with predictions of his potential performance at colleges in which he is interested.

The College Entrance Examination Board recently has begun offering an information system, the Comparative Guidance and Placement Program (CGP), specifically for use in community colleges. The ACT and CGP programs are alternatives appropriate for students in academic curriculums in community colleges. A new guidance-oriented information system, the Career Planning Profile (CPP), is currently under development by ACT for use by students in vocational-technical curriculums. The CPP and CGP programs are alternatives for students in these curriculums.

Thus, for the past decade, we have been witnessing a continuing reorientation of services offered at the postsecondary level by the major testing organizations (24). The present trend will undoubtedly continue, and Bayesian statistics can, I think, make an important contribution in this new setting (17).

The Bayesian method is unique in providing a formal mechanism for combining observational information with *prior* information or beliefs to provide *posterior*, or after the sample, probability distributions for parameters of interest such as student abilities, institutional mean values, or regression coefficients relating performance criteria to test scores. A typical Bayesian statement made after observing a small random sample of persons would be of the following form: the probability is .95 that the mean ACT English score of examinees from Iowa in the year 1969 lies between 20.4 and 23.2. The length of such a *credibility interval* would depend largely on the number of observations in the sample.

The posterior probability distribution is interpreted by Bayesians as a formal numerical representation of the state of knowledge about the parameter of interest. It literally carries all of the available information about the parameter. Certain characteristics of this posterior Bayes distribution are of particular interest. For example, such measures of central tendency as the mean, the median, and the mode are useful as general descriptors, the mode being the most probable value of the parameter. The reciprocal of the variance of the posterior distribution is a measure of the *precision* of available information.

The heart of the Bayesian method is Bayes' theorem which says that, given the data, the posterior distribution of the parameter is proportional to the product of the distribution of the data, given the parameter, and the prior (or before the sample) distribution of the parameter. The first of these distributions is what is often called the model distribution and is simply that used in classical forms of parametric inference. Bayes' theorem itself is a straightforward application of the basic theorem of conditional probability and hence enjoys general acceptance. In effect, Bayes' theorem adds sample information to prior information to provide a formal representation of posterior information. The Bayesian method may thus justifiably be thought of as a formal system of information accumulation.

In many simple applications Bayesian credibility interval statements either coincide numerically with classical confidence interval statements or differ only by trivial amounts. The two kinds of interval statements, however, have quite different meanings. The classical statement is "the probability is .95 that the obtained confidence interval will cover the true mean." This is a statement about the interval, not the mean. The Bayesian statement is "the probability is .95 that the true mean lies in the specified credibility interval." The Bayesian statement is a direct statement about the mean; many people find it preferable.

The price one pays for the elegance of the Bayesian analysis is the need for specifying a prior Bayes distribution summarizing prior information or beliefs. There is controversy on this point because, first, some people do not wish to interpret probabilities as degrees of belief, but only as relative frequencies as in classical theory and, second, even accepting a belief interpretation for probabilities, there still remains a very real problem of just how to quantify these beliefs. The latter problem is particularly acute because, in any important study, experts will disagree on the evaluation of prior information. Indeed, the purpose of the study is typically to resolve such disagreements.

Bayesian Considerations

In 1963 a major paper by Edwards, Lindman, and Savage (8) describing Bayesian methods appeared in the *Psychological Review*. This paper described the Bayesian method as an explication of a theory of personal probabilities with which the names of Ramsey (19), de Finetti (7), and Savage (21) are most prominently associated. The impact of this paper was enhanced by the enormous popularity that Bayesian methods were enjoying in business applications, primarily as a result of the efforts of Schlaifer (22).

The Bayesian personal probability method is described as resting on two foundational supports. The first of these, developed in the *Review* paper, is a theorem showing that if each investigator uses a reasonable prior distribution, all posterior distributions will eventually converge and we will thus have *stable* estimation. Thus, the Bayesian method is shown to have the requisite property of eventually resolving prior differences of opinion.

The second support for the theory is based on an argument due to de Finetti and formalized in a theorem by Savage (21). In essence the theorem says that if you wish to be sure of behaving in a logically consistent or *coherent* manner in any decision situation, then you must effectively behave as if you had a prior distribution and you must effectively use Bayes' theorem. An implication of Savage's theorem is that if you behave in a non-Bayesian way in a betting situation, your opponent can specify a sequence of bets that would appear favorable to you and that would, in the long run, almost certainly lead to a loss by you. One might expect these arguments to be compelling, for who would choose to bear both the professional scorn and the economic ruin that logical inconsistency promises to bring.

Many papers have also appeared showing that well accepted principles of classical inference can lead to very unsatisfactory results (1, 4). For example (18), the usual classical unbiased estimate of a between-group variance component can be negative even though a variance component must, by definition, be non-negative. In contrast, the Bayesian estimate is always non-negative. Despite this, the Bayesian method did not receive on-the-spot acceptance because of a perceived weakness involving the selection of the prior distribution. According to the personal probability theory, each investigator constructs his own prior distribution by means of a self-interrogation or introspection of how he would bet on various possible values of the parameter. No attempt is made to attain any sort of preexperiment consensus among investigators; rather, great reliance is placed on the principle of stable estimation.

The usual objection raised to personal probabilities is that it is the antithesis of science to let each experimenter select his own prior distribution. Somehow, it is thought, the prior information must depend on prior data. This is very difficult, however, because prior information is typically fragmented and the evaluation of it is subject to individual interpretation and bias.

It also seems evident that, while the business entrepreneur need convince only himself of the reasonableness of his action, the scientist is typically trying to convince someone else—a journal editor, a research grant committee, or the readership that a conference such as this one provides. It seems to me this necessitates, in scientific publications, that one of two things must be done. Either the prior distribution must be as well justified as anything else in the study or, for argumentative purposes, the scientist must present a parallel analysis showing that even with a prior distribution that others might specify, the results of the present experiment support his contentions.

The technique I now wish to discuss makes it possible to construct a prior distribution from the data at hand and thus largely to depersonalize personal probabilities. This technique can be used whenever inferences are made simultaneously about a large number of persons, schools, or other experimental units—for example, in estimating the true scores (that is, expected scores) of members of a well-defined group of examinees. We know that the observed score for a person has an error distribution centered at his true score. But since we treat our examinees as having come from a population of potential examinees, we also have a distribution of (unobservable) true scores. Thus, we have the well-known model II, the variance components or random effects model, which has been studied along classical lines by many statisticians including Cornfield and Tukey (5). The model has been used in a semi-Bayesian way to estimate means by Robbins (20) and by Stein (23). Earlier still, this model was used to estimate means in educational work by Kelley (13). Recently Bayesian analyses for the estimation of means with this model have been provided by Box and Tiao (2) and by Lindley (14) and applied in the field of public health by Cornfield (3). A comparison of some Bayesian and classical methods has been done by Novick, Jackson, and Thayer (18).

The Kelley regression estimate of true score given observed score has a form that closely approximates other model II solutions. That estimate is just a weighted average of the person's observed score and the mean observed score in the population, the weights being, respectively, the reliability of the test and one minus the reliability. Thus the regression estimate of true score depends not only on the direct observations available on the particular person but also on the indirect or *collateral* information gained from all other observations in the specified group.

This regression estimate makes sense. If we have an unreliable measurement on any person, a heavy weight is given to the mean value of the population of which he is a member and the estimate is regressed back nearly to that value. If our measurement is very reliable, it gives little weight to this population value and there is very little regression. In intermediate cases there is only partial regression to the overall mean. Kelley (13) showed that the overall mean squared error is substantially reduced by using this procedure when the reliability itself is low or moderate.

The various Bayesian and semi-Bayesian approaches to this problem yield results that are very similar to those obtained by Kelley. Robbins (20) captured the spirit of what was being done when he preempted the name *empirical Bayes* for his procedure. In effect, what is being done here is to use the collateral observations to estimate the parameters of the prior distribution for each person and then to use the direct observations to get the posterior distribution. Robbins' procedure differs from the full Bayesian model II analysis in that he uses a classical method to estimate the parameters of the prior distribution for the Bayesian analysis, while the full Bayesian analysis also does this in a Bayesian way. My own feeling is that the new Bayesian procedures are as empirical as Robbins' procedure, possibly more so. They are certainly more illuminating theoretically, and only these new procedures provide a formal method for combining both prior and collateral information.

A third foundational support for Bayesian work—and particularly for Bayesian model II analysis—is contained in a theorem, due to de Finetti (7) and generalized by Hewitt and Savage (10). If our prior information about the various persons is identical, then we must have what de Finetti calls a symmetric or *exchangeable* prior distribution for the person parameters. The de Finetti-Hewitt-Savage theorem states that any exchangeable prior distribution is equivalent to a prior distribution obtained under the assumption that the persons were randomly sampled from some population, and hence that model II is applicable. The strength of this theorem now seems very great. It means that a model II analysis will *typically* be preferable to a model I, that is, fixed effects analysis (14).

Despite our well-displayed fondness for the Bayesian model II estimation of means, we must acknowledge that there can be a problem. It may add to overall efficiency to reduce our estimate of a person's true score because we identify him with some population that has a lower mean true score, but it may not appear fair. Suppose, in a selection situation, that one person has his score lowered by this regression to the population mean and a second person from a population with a higher mean true score has his score raised. Suppose further that this results in an inversion in the ordering of the reported scores and that, as a result, the second person is selected for college admission and the first is not. We would certainly be hard put to convince the first examinee, his parents, *and his lawyer* that he had been treated fairly.

We do not mean to suggest that model II cannot be used in a selection situation, only that to do so fairly may require a much more careful selection procedure; one—for example, that considers in a full decision-theoretic analysis the differential utility of accepting persons from the different groups. The important point, though, is that the whole situation changes when the student becomes the decision maker, that is, when we are considering a guidance rather than a selection situation. The decision of what to do with this information then falls to the student. He may, for example, want to modify our estimate, using information available to him but not to us.

Actually, the above discussion is largely academic with a test like the SAT, which is very long and reports only two scores and therefore has high subtest reliability. The regression estimates of true score will then differ little from the observed score. In multi-scale batteries of short subtests the effect on subtest scores will be more pronounced. In such situations one might find merit in reporting the Bayesian multiple regression estimate of each true score given all of the observed scores. This approach has been suggested by Cronbach and Furby (6) for the estimation of change scores. Since only a single overall population is identified, there will be no unfairness to any individual. When the intercorrelations of the subtest scores are more than trivial, this can result in a substantial increase in the reliability of each subtest.

When used to estimate institutional parameters or regression coefficients, in either a guidance or a selection context, the model II estimates are also not subject to any unfairness criticism. This application is important because by using prior and collateral information in a Bayesian analysis we can typically obtain any specified degree of precision with a smaller sample size than a model I analysis would require. It really makes no sense to estimate each institutional parameter, or for that matter to do every validity study, as if we were starting from a state of ignorance.

The most immediately important application of the Bayesian model II analysis, in my judgment, is to the estimation of regression parameters. Each of the guidance-oriented testing programs mentioned earlier incorporates predictions of academic performance as an important piece of information to be supplied to the student. The growth in the number and diversity of programs at the community college level and the relative smallness of individual programs suggest that often we shall not have enough data on a particular curriculum within a particular college to estimate the partial regression weights with satisfactory accuracy. Analyses that we have done on data from each of the three guidance testing programs confirm this expectation. The problem will become even more acute as we sharpen our focus on post-training criteria and are then inevitably faced with drastically reduced sample sizes.

What we will need to do is recognize that in carefully specified groupings of community colleges, for example, regression coefficients for a particular curriculum do not differ too greatly across colleges. We can expect some differences in the regression weights because of minor differences in curriculum content and grading standards, but a great deal of similarity can be expected.

Recently Professor D. V. Lindley of University College, London, has supplied us with a full Bayesian model II analysis for regression in m colleges. The result of this analysis in the single predictor case is to regress the regression weight for each college towards the average of the regression weights across colleges. Here the amount of regression depends largely on the true variance of the regression weights across colleges and on the sample size within the particular college. According to statistical theory, the Bayesian estimates of the regression weights should, on the average, be more accurate than the usual model I estimates. We have now completed the programming of Lindley's very complex solution to this problem and have applied the technique extensively to the estimation of regression parameters obtained from one testing program. We have done this for both simple linear regression and for multiple regression.

Table 1 gives the results of one such analysis. The usual least squares estimates of model I are given in the first column. Notice that two of these estimates are negative. Neither I nor any person I have consulted really believes that the true values are negative. In the second column the estimates obtained from Lindley's model II Bayesian analysis are given. These values certainly more nearly correspond with what we think the true state of affairs to be.

In order to check the reasonableness of our Bayesian solution, we have also developed a classical model II analysis (12, 11). The third column of Table 1 gives the values obtained from this analysis. The relative closeness of the solutions in columns 2 and 3, and their substantial difference from the solution in the first column, suggest to us that the Bayesian solution is both accurate and useful. Recent data analyses that we have done suggest that predictions based on the

Table 1

Comparison of Three Estimates of Regression Coefficients
Comparative Guidance Program—Education Curriculum
Regression of GPA *on Vocabulary Score*

College No.	*Least Squares Estimates*	*Bayesian*	*Classical Model II*	*College No.*	*Least Squares Estimates*	*Bayesian*	*Classical Model II*
1	2.2	2.9	2.7	11	1.5	2.7	2.2
2	−1.6	2.0	0.4	12	3.1	3.1	3.1
3	5.1	3.6	4.0	13	2.6	3.0	2.7
4	4.9	3.9	4.4	14	3.4	3.1	3.4
5	2.6	3.0	2.8	15	3.8	3.4	3.5
6	−0.1	2.2	1.7	16	2.2	2.8	2.6
7	9.3	4.4	6.3	17	1.1	2.4	1.7
8	3.4	3.2	3.3	18	3.9	3.6	3.7
9	3.7	3.4	3.5	19	4.0	3.5	3.8
10	0.1	1.9	1.1	20	4.7	3.9	4.3
				21	5.9	4.0	5.0

Acknowledgment is made to Educational Testing Service for making data available for this analysis.

ACT Test will similarly benefit from a Bayesian treatment. I should also mention that an empirical Bayes procedure for this problem (15) has also recently been published, but we have not yet completed our study of this work.

The assumptions upon which the Lindley derivation is based require that this kind of analysis be done by a Bayesian statistician only in close collaboration with an educational specialist. The grouping of colleges into homogeneous groups in order to satisfy the exchangeability assumption may be very important. We have high expectation that empirical work will show that when the Bayesian method is carefully applied, it will yield very meaningful improvements in prediction over the classical model I analysis. If this is true, Professor Lindley's work will prove to be a major contribution to guidance technology and more generally to the development and use of educational information systems.

REFERENCES

1. Bock, R. D. and Wood, R. Test theory. *Annual Review of Psychology*, 1971, *22*.
2. Box, G. E. P. and Tiao, G. C. Bayesian estimation of means for the random effect model. *Journal of the American Statistical Association*, 1968, *63*, 174–181.
3. Cornfield, J. The Bayesian outlook and its application (with discussion). *Biometrics*, 1969, *26*, 617–658.
4. Cornfield, J. The frequency theory of probability, Bayes' theorem and sequential clinical trials. In D. L. Meyer and R. O. Collier Jr. (Eds.) *Bayesian statistics*. 9th Annual Phi Delta Kappa Symposium on Educational Research. Itasca, Illinois: Peacock Publishers, Inc., 1970.
5. Cornfield, J. and Tukey, J. W. Average value of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, *27*, 907–949.
6. Cronbach, L. J. and Furby, L. How should we measure "change"—or should we? *Psychological Bulletin*, 1970, *74*, 68–80.

7. de Finetti, B. Foresight: Its logical laws, its subjective sources. *Annales de l'Institut Henri Poincaré*, 1937, *Vol. VII*, (Reprinted in Kyburg, H. E. Jr. and Smokler, H. E. *Studies in subjective probability*. New York: Wiley, 1964.)

8. Edwards, W., Lindman, H., and Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review*, 1963, *70*, 193–242.

9. Harcleroad, F. F. (Ed.) *Issues of the seventies*. San Francisco: Jossey-Bass, 1970.

10. Hewitt, E. and Savage, L. J. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 1955, *80*, 470–501.

11. Jackson, P. H. The estimation of many parameters—some simple approximations. *ACT Research Report* (in press). Iowa City, Iowa: American College Testing Program, 1971.

12. Jackson, P. H., Novick, M. R., and Thayer, D. T. Bayesian inference and the classical test theory model. II. Validity and prediction. *ETS Research Bulletin* 70–32. Princeton, N. J.: Educational Testing Service, 1970.

13. Kelley, T. L. *Interpretation of educational measurements*. Yonkers-on-Hudson, New York: World Book, 1927.

14. Lindley, D. V. The estimation of many parameters. *Proceedings of the Waterloo Conference on the Foundations of Statistics* (in press).

15. Martz, H. F. Jr. and Krutchkoff, R. G. Empirical Bayes estimators in a multiple regression model. *Biometrika*, 1969, *56*, 367–374.

16. Munday, L. A. and Rever, P. R. Perspectives on open admissions. Concluding chapter in P. R. Rever, (Ed.) ACT Monograph Four: *Open admissions and equal access*. Iowa City, Iowa: American College Testing Program (in press).

17. Novick, M. R. and Jackson, P. H. Bayesian guidance technology. *Review of Educational Research*, 1970, *40* (No. 4), 459–494.

18. Novick, M. R., Jackson, P. H., and Thayer, D. T. Bayesian estimation and the classical test theory model: Reliability and true scores. *Psychometrika* (in press).

19. Ramsey, F. D. *The foundations of mathematics and other logical essays.* London: Kegan, 1963.

20. Robbins, H. An empirical Bayes approach to statistics. In J. Neyman, (Ed.) *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. I.* Theory and statistics. 1954/55, 157–163.

21. Savage, L. J. *The foundations of statistics.* New York: Wiley, 1954.

22. Schlaifer, R. *Probability and statistics for business decisions.* New York: McGraw-Hill, 1959.

23. Stein, C. M. Confidence sets for the mean of a multivariate normal distribution (with discussion). *Journal of the Royal Statistical Society,* 1962, *24*, 265–296.

24. Turnbull, W. W. Relevance in testing. *Science*, 1968, *160*, 1424–1429.

Discussion

JOHN W. TUKEY
Princeton University

It seems to me I could comment on this paper from various directions, and I will say a word or two from perhaps two directions. I call your attention to one of the latter statements when we were told that "Not only do we have a Bayesian solution, but we have a classical solution, and they agree fairly well." and then we are told by Dr. Novick that "the relative closeness of their solutions suggest to me that the Bayesian solution is both accurate and useful."

I think it would be interesting to consider what Dr. Novick would say if somebody rose up to say that the paper suggested to him that the *classical* solution was both accurate and useful. If you look hard at those numbers, you will find that the changes from the least square solution are about in the ratio of 100 to 55 with a few exceptions. I think it might be interesting some time to inquire into the exceptions, though I don't think it is important here. What I think is important to say is this: Given the data from which this example was drawn, it seems to me perfectly possible to ask of that same data which of these two approaches seems to be working better, and by what other factors would it be good to multiply the changes that each of them implies, in order to get as good a result as you can by this type of adjustment.

I am sure this factor will not be zero. I am sure this is a good sort of adjustment. I have no burning principle that tells me whether the amount of adjustment from the classical model II or the amount of adjustment from the Bayes is going to give the better results, but since we have computers and computations often at very reasonable cost, we could perfectly well do a leave-out-one type of validation study here in which we leave each student alone out of the computation, each one in turn, go through and do everything over, and then use the two prediction formulas to see how that student should have come out.

If we do this for all students, and average, this is an honest cross-validation procedure for the two methods of setting regression weights, and we ought to be able to tell whether the difference between these is a matter of one percent or a tenth of a percent or maybe three percent. Maybe somebody has tried this and could give us an idea how many percent it should be.

In other words, I suspect the difference is small. I am pretty sure both of these methods are better than the direct least squares approach, and I think it is a perfectly answerable question to make some comparisons between each of these and use, say, one and a half times the classical change or three-quarters of the Bayes. If that's better, I would be prepared to use it.

In passing I think one should notice that the words "model II" are, from my point of view, not being used in quite their usual sense. I don't think it is confusing or dangerous, but if you look at model II in a textbook you won't get quite this.

Let me turn to the major part of the paper which, in the good sense of the word propaganda, let me call Bayesian propaganda. I am still neutral to the Bayesian question, which implies that I am inclined to believe there are situations where it will help but also that I am inclined to believe there are others where it won't. But I think there are some comments to be made about some specifics.

On the same point about which I quoted earlier, it is stated that "the assumptions require that this kind of analysis be done by a Bayesian statistician only in close collaboration with an education specialist."

If this is really true, I think we ought perhaps to hold this as a practical weakness of the situation. If the classical model II operates the way I would expect in terms of minimum mean square error, it is going to help us whether or not we have been able to put the colleges into perfectly homogeneous groups. Groups that have some real differences will do us some good.

At an earlier stage there was an assertion about the uniqueness of the Bayesian technique as a way of combining collateral and prior information. I guess it seems to me the fact that one is willing to lean on a classical model II method implies that the classical model II method must in fact be providing the same sort of combination.

And finally, on the other side, I would say if Bayes techniques helped to bring forward such approaches and such techniques as we have heard today, then they may be serving a very useful purpose whether or not the final decision is to use Bayes techniques or classical techniques.

One last point that you would have had no chance to see. With respect to the discussion of model II, I find it interesting to note that the preliminary version of the paper read that doing this "may not be fair to the individual," whereas the final version said "may not appear fair." That is an interesting difference. And I am not sure just where we stand on this.

DR. MELVIN NOVICK: Well, after seven years of discussion of topics like this with John, apparently we have come somewhat closer in agreement than we once were. I want to read something into the record here:

> The relative closeness of the solutions in columns two and three and their substantial difference from the solution in column one suggest to us that the classical model II solution is both accurate and useful.

I fully subscribe to that.

Actually, I thought the statement I made would be more acceptable because it sort of suggests that I am validating the Bayesian solution on the classical solution, but if John wants to justify the classical solution because of its closeness to the Bayesian one, that's fine.

There is a larger area of agreement between John and me. I am perfectly willing to use empirical Bayes, the Stein procedure, and classical model II, particularly when I am doing data analytic kinds of things, but I have the feeling, which I can't document now, that when we talk about educational information systems—and we are talking about educational information systems here—that the Bayesian approach will be the preferable route to go. When we talk about educational systems, we are talking about a situation where we are going to have an educational specialist working in close collaboration with a statistician, and presumably most statisticians in a few years will know something about Bayes. There is just one small point of puzzlement on my part. I see how the Stein, Robbins, or the classical model II approaches incorporate what I have called "collateral information," but the Bayesian method gives a formalism for incorporating quite different kinds of information.

I have just read the abstract of the paper by Martz and Krutchkoff on the empirical Bayes approach to this problem. They say they are getting a substantial improvement in mean squared error using an empirical Bayes approach. Now that makes me feel much more confident than I was before seeing that paper, and if I had to bet on our Bayesian thing working, I would bet quite a bit more boldly than I would have a month ago.

Now I think I know how to use a Bayesian analysis to incorporate prior information like that. I don't know how to do that in the classical context and if John does, I wish he'd tell me.

I believe that I have said all that I can say for the present on the fairness question. When any of the methods being discussed here, classical or Bayesian, is used carefully, there should be no problem; but these techniques can be misused, as can all statistical techniques. It is important that this danger be given wide publicity so that due caution can be observed; but in the applications that I have discussed there will be at most an appearance of unfairness.

Temporal Changes in Treatment-effect Correlations: A Quasi-experimental Model for Institutional Records and Longitudinal Studies[1]

DONALD T. CAMPBELL
Northwestern University

This paper has two general goals. The first is to present some quasi-experimental designs particularly appropriate to the utilization of educational records and data from longitudinal or multiwave panel studies. The second, and perhaps more important in the long run, is to search for experimental designs appropriate for situations in which people volunteer for experimental treatments. At the present time there are no designs available that will adequately distinguish between treatment effects and cosymptoms of the selection differences that volunteering produces. Yet the "experimenting society" of the future (4) must also be a voluntaristic one, avoiding the coercive control implied in randomized assignment to treatments (14). We are each of us convinced, in terms of our own experience, that treatments we have volunteered for—the jobs, wives, curriculums, psychotherapies, and so on, that we have chosen—have changed us. While part of this may be a causal-perceptual illusion akin to the statistical regression artifact, surely not all of it is. Eventually the ponderous processes of science should also be able to see what is thus visible to the naked eye.

Consider a study in which attributes of children (such as vocabulary, mathematical skills, problem-solving ingenuity, and so on) are repeatedly measured on the same children over a substantial number of years, and in which specific experiences not uniformly shared (such as courses in new-math, Head Start, Follow Through) are recorded. While in a true experiment these experiences, these potential change-agents, can be assigned at random to a subsample and withheld from

---

[1] This paper was supported in part by funds from National Science Foundation Grant #GS1309X.

an equivalent group, in our situation this has not been possible. Instead, selection and treatment are confounded; those getting the treatment differ systematically even before the treatment.

The usual approach to such initial differences is to attempt to adjust them away. Not only have such adjustments proven inadequate; they have, as a by-product of the chronic underadjustment, produced results with systematic biases. For that class of treatments given to those who need them least (such as accelerated tracks, honors courses, and university education), these may often seem benign errors, merely exaggerating the efficacy of treatments we know in our hearts to be good. But for a treatment we give to those who need it most (such as remedial reading or Head Start), the bias is in the direction of making the treatment look harmful, and thus of underestimating or swamping any true effects. It seems to me certain that the Westinghouse-Ohio University evaluation of Head Start (8) contained such a bias, a tragic error when one considers that this study was used to justify the destruction of the Head Start program, and was probably the most politically influential statistical evaluation ever done up to that time (6). Not only do "matching," *ex post facto* analysis, and "control" by partial correlation produce such regression artifacts (for example, 21, 1), but so does analysis of covariance (18, 23, 6, 9).

Living with Pretreatment Differences Rather Than Adjusting Them Away

One basic recommendation in the present paper is that we give up trying to adjust away pretreatment differences. Rather, we should live with them, use them as a base line, and demand that an effective treatment significantly modify that difference.

There are numerous statistical symptoms of an experimental treatment effect (5). The common ones of mean differences or differences in change scores must be ruled out for growth data on children because pretreatment differences almost certainly imply preexisting differences in growth rates as well, as illustrated in Figure 1. Such divergent growth rates no doubt occur within groups as well as between groups, the increased separation of means being accompanied by increased variability of groups, in what we can call the "fan-spread hypothesis" (2). Indices such as t or F, which express mean differences relative to

Figure 1

Experimental effects as changes in the treatment-effect correlation
In these hypothetical "true" experiments,
the correlation for the pretest is zero.

PRETEST

Treatment	1	2	3	4	5	6	7	8	9	10
4	5	10	20	10	5					
3	5	10	20	10	5					
2	5	10	20	10	5					
1	5	10	20	10	5					
0	5	10	20	10	5					
Total	25	50	100	50	25					

POSTTEST

Treatment	1	2	3	4	5	6	7	8	9	10
4					5	10	20	10	5	
3				5	10	20	10	5		
2			5	10	20	10	5			
1		5	10	20	10	5				
0	5	10	20	10	5					
Total	5	15	35	45	50	45	35	15	5	

PRETEST

Treatment	1	2	3	4	5	6	7	8	9	10
1	5	10	20	10	5					
0	5	10	20	10	5					
Total	10	20	40	20	10					

POSTTEST

Treatment	1	2	3	4	5	6	7	8	9	10
1			5	10	20	10	5			
0	5	10	20	10	5					
Total	5	10	25	20	25	10	5			

variability avoid the difficulty. Thus, the recommendation becomes that of computing the pretreatment t between experimental and control groups, and comparing the post-treatment t with it, an experimental effect being shown as a significant difference in t's, rather than a posttest t significantly different from zero.

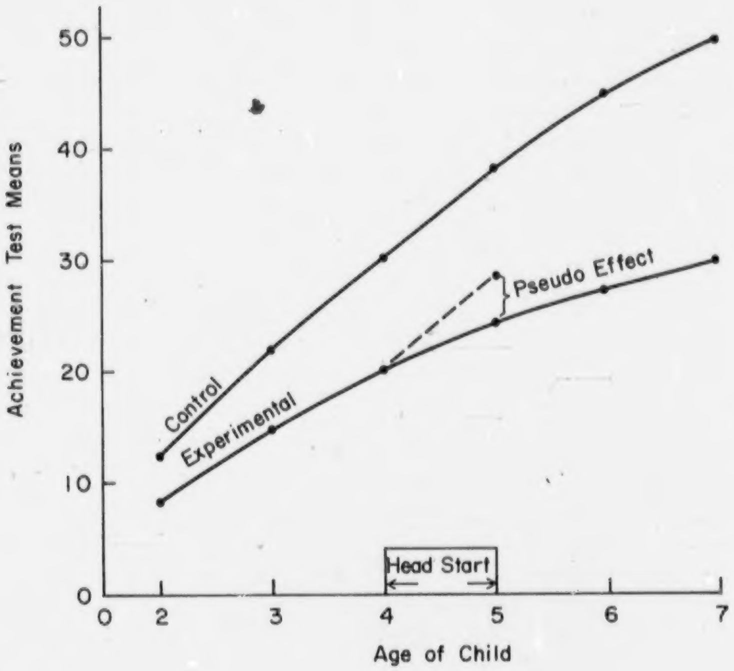
In what follows, instead of t or F, an r between the treatment taken as a variable and the dependent variable will be used. This r is also an expression of mean differences relative to variability. For example, a biserial r is computed from the same ingredients as are found in a t. The preference for r over t or F is arbitrary, but it has the advantage of being descriptive of the strength of relationship independently of the number of observations employed. More importantly, r makes conceptual contact with the correlation-causation problem as explored in the lagging of time-series correlations (12, 26) and in the cross-lagged panel correlation (3, 22, 25, 24, 16, 17).

In case r seems an unusual measure of an experimental effect, Figure 2 is provided. The top scatter diagrams illustrate pretest and posttest distributions for an experiment involving four degrees of the treatment variable, plus a control condition. For the pretest, due to random assignment (from sets of five matched pretest scores in this case), all groups have the same mean and standard deviation. The correlation between treatment levels (O = control, 1, 2, 3, 4) and the pretest scores is thus zero. For the posttest, r has acquired a high positive value. If the effects had been nonordinal, one would need to use a curvilinear or nonordinal measure of relationship, such as eta, or a contingency coefficient. The effect, of course, might be negative rather than positive, but in any case, in a true experiment, the correlation would start at zero for the pretest, and goes on the posttest to some value positive or negative, significantly different from zero, if there were a treatment effect. In the lower half of Figure 2 is portrayed the more usual situation in which there is only one experimental group and one control group. Here, too, one can use the correlation concept. The biserial r (and the t) start at zero for the pretest, move to a substantial positive value for the posttest.

For quasi-experiments where the correlation does not start at zero, it is here proposed that we give up as misleading all statistical efforts to adjust it back to zero (by matching or covariance, and so on) and instead demand that a treatment effect show itself as a significant change in the treatment-effect correlation, a significant increase or decrease.

Figure 2

Illustration of the pseudo effects possible if the differential growth rates associated with initial mean differences are disregarded



Temporal Erosion

But experimental treatments are not the only processes that change treatment-effect correlations. All relationships tend to weaken with time, a process we have previously designated as "temporal attenuation" (25), but to avoid confusion with ordinary reliability processes we now call "temporal erosion" (16, 17).

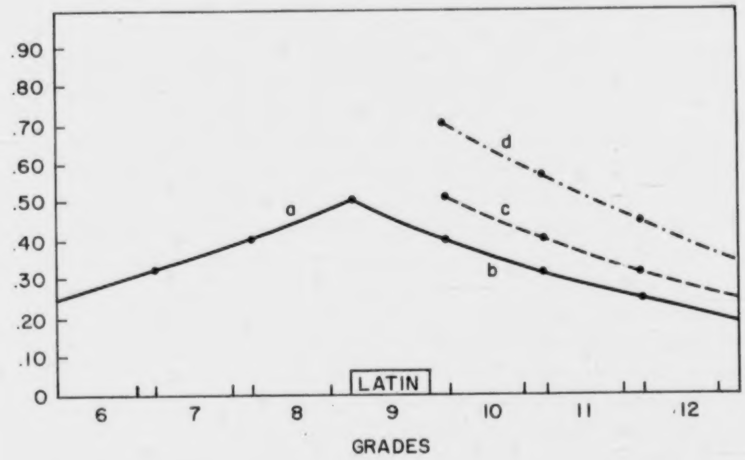
Let us first consider a series of repeated measures in the middle of which a treatment has been given. Annual September English vocabulary scores and a ninth grade course in Latin can be used for illustration. The biserial correlation of vocabulary with the presence or absence of Latin is computed. In Figure 3, a no-effect outcome and an incremental effect of Latin are plotted.

Figure 3 presumes that *all relationships erode in time*, and that *erosion rate is constant over equal time periods*. In the graphed values, the erosion rate is .80. (The no-effect values are .50, .40, .32, .256, and .2048. The effect values of line *d* are .70, .56, .448, and .3584.) The assumption of constant erosion rate means the slopes would appear linear when plotted in logarithms.

The erosion rate for a correlation is presumably a product of erosion characteristics of both variables. Since the "measure" called Taking Latin occurs only once, we have no additional grounds for estimating its rate. (The erosion rate for Taking Latin as a symptom or measure is also to be distinguished from the dissipation rate for the real effects of Taking Latin, if any. Figure 3 assumes that the composite of Latin as symptom, Latin effect, and English vocabulary as symptom attenuates at .80.) The correlations among the vocabulary measures provide bases for evaluating its rate, and, for it, the validity of assumptions A and B. The matrix of such relationships should be "proximally autocorrelated" (25) or of a "superdiagonal" type (20) or a quasi-simplex (10, 13) in form. That is, the correlations between adjacent time periods should be higher than those spanning two periods, and these higher than those spanning three periods, and so on. The "slope" of these correlations away from the diagonal is not apt to be technically what Guttman has called a simplex, forming a uniform pattern if unities are placed in the diagonal, but instead will have implicit values in the diagonal lower than 1.00, as in Table 1, and will presumably correspond to a first order autoregressive function or a Markov process (15). This corresponds to a uniform rate of erosion, a uniform rate of degrading the relationship by substitution of error or mismatching persons. (If there has been an effect of Latin, this might affect the intercorrelations of the vocabulary tests. We should accumulate experience from true experiments on this. Is the test-retest correlation higher in the experimental or control group?) If there are grounds for ascertaining erosion rates separably for each variable, the erosion rate for the correlation might be assumed to be the geometric mean of the two, in analogue to the correction for attenuation in reliability, and on the assumption of homogeneous erosion of all components within a given variable.

Figure 3

Biserial correlation of annual September vocabulary tests with taking Latin
Line b *is a clear-cut case of no-effect,*
line d *a clear-cut case of incremental effect.*



Locating the "Correlation Peak" for the No-effect Condition

With this background, we can begin to consider the problematics of any specific instance. The pretreatment correlations with Latin are due to the fact that taking Latin is a symptom of common determinants that also produce high English vocabulary scores. The peak in this correlation comes at the point of simultaneous "measurement." If "intention to take Latin at the first opportunity, that is, in the ninth grade" were measured in the sixth grade, the correlation of vocabulary and this "Sixth Grade Intention" would peak in the sixth grade. Note in Figure 3 that we have peaked the no-effect curve at the beginning rather than at the middle of the year of Latin. It should be peaked at the point where the decision was made, at "registration" if Latin is optional. What if Latin is an obligatory part of a track system and all pupils on one track receive it? Then presumably the peak is at the last point of actual or potential revision of track membership prior to Latin. Note in this case that it is of help to have the several pretest measures. If the tracks were fixed when pupils entered junior high in the seventh grade, then the correlations should peak at 7 tapering off through 8, 9, and by extrapolation, 10 and 11.

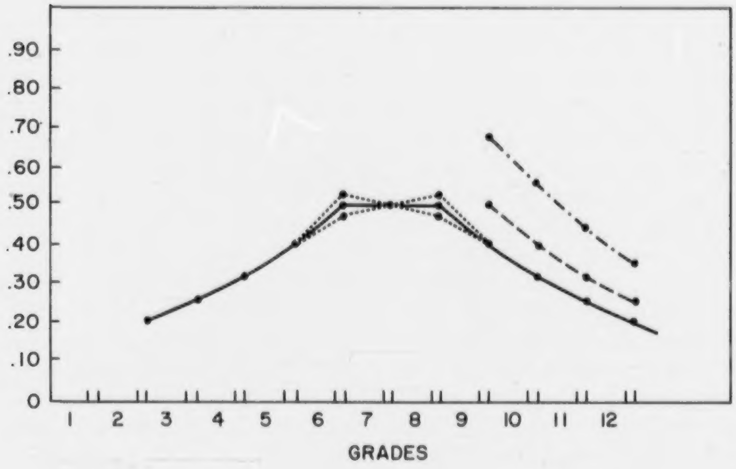
The judgment as to when the decision or determination was made will be important in interpreting weak effects, such as outcomes lying between *c* and *b* in Figure 1. The coarse grain of the measurement series (the wide spacing of measurements) will increase the ambiguity. Almost certainly, the decision, and hence the peak, will occur prior to the treatment, with how much prior being the question. Thus, an outcome like *d* will stand as an unequivocal effect whatever decision point and whatever temporal erosion rate one assumes. An outcome like *c*, even though the correlation after is the same as before, is usually also symptomatic of a positive effect for reasonable fixings of decision point and temporal erosion rate (but see below).

Types of decision processes vary in their temporal location and sharpness of focus. In Figure 3, we have assumed a voluntary choice of courses made at the beginning of the term, and maximally symptomatic of the pupils at that moment. At another extreme, the assignments would be decided by the high school staff at the beginning of the term, but based upon the pupils' grades of the prior year. In this case, the decision point, and the correlation peak under the no-effect case, lies sometime in that prior year, depending upon the weighting given to various semesters and the intercorrelation of grades from semester to semester. Not only is the peak earlier, but it is also less focused, more spread out. Intermediate and more characteristic conditions would include setting prior-performance prerequisites for Latin or heavy influence of teacher's advice, the latter being based upon prior performance, and so on. All of these move back and spread out the time in the pupil's career maximally symptomatized by the decision to take Latin.

In a situation in which pupils can freely drop or transfer out of Latin, and in which considerable numbers do, staying in Latin becomes a selective diagnostic of ability and interest, and so on, which has its time of maximal symptomicity toward the end of the Latin course. If the situation were completely fluid, with each day of Latin requiring a new commitment made without cost in either direction, then the symptom of attending the last day of Latin would have its peak at the end of the treatment. Probably all reasonable analyses

Figure 4

Figure 3 modified for revisable seventh grade tracking into Latin



would show that decisions in or out are greatly increased in difficulty and rarity once the term has begun, and that later-term drops are due to the symptom-load of early term performance; hence, no reasonable model would put the correlation peak later than the middle of the Latin treatment since a middle placement jeopardizes the interpretation of outcome *c* in Figure 3, but not outcome *d*.

More likely than complete fluidity of decision, or homogeneity of redecision in time, is a stepwise process of major decisions and reluctant revisions. These would create erosion patterns with plateaus in them. Figure 4 illustrates a case in which all those in the top junior high track take Latin, the tracking decision being made at entry to seventh grade, but with minor revisions and transfers made each year.

Getting into a track at the beginning of the seventh grade is much easier than changing in or out in the eighth or ninth grade. There results some kind of correlation plateau in the seventh-to-ninth region. Whether this tilts up toward ninth or up toward seventh depends upon the relative strengths of the selective factors. A procedure which let no more in, but continually purified by elimination the group selected at seventh, might correlate higher at the end of the process, at grade nine.

A sharp focused peak will result from assignment to Latin on the basis of a test, given on a specific date, which correlates with the English vocabulary test. The date of that test will be the peak. The sharpest peaking would result from using the English vocabulary test itself as the basis of assignment to Latin. This would produce a peak at the level of 1.00, making it impossible to achieve an unequivocal evidence of effect, that is, a post-treatment *r* higher than the peak. The lower the pretest-treatment correlations and the lower the presumed peak, the clearer the experimental inference. A decision base which correlates zero with English vocabulary would be as good as randomization, with no peak, all pretest values and erosion slopes flat at zero.

Hidden peaks are a threat to this analysis. Since the sharp peaked decisions will occur before the onset of the treatment, an immediate pretest such as assumed in Figure 3 will protect against a hidden peak masquerading as a treatment effect. But if the nearest pretest were in June of the previous year, and if the decision were made on a September language aptitude test at the beginning of the ninth grade, then the failure to ascertain this peak might lead to an underestimation of the no-effect level for posttest values.

ESTIMATING EROSION RATES AND INTERCEPTS

Before further wallowing in potential equivocalities, it should perhaps be announced that problems of both peak location and erosion rates are probably exaggerated in the .80 rate used in Figure 1. Analyses of the data from the big ETS STEP-SCAT longitudinal study, covering grades seven through eleven, show biannual erosion rates of .95 to be typical, .90 to be minimal (11, 16). (Perhaps these should be called nonerosion rates, since 1.00 would mean no erosion at all.) Such high rates mean that the peaks are only slightly higher than the other values, that equivocalities in the location of the peaks, or in estimating rates, create only narrow ranges of equivocality in estimating the no-effect expectation for post-treatment values.

The simplest rate assumption is uniformity in time, both forward and back from the decision point. In a limited way, the assumption can be checked in the pretest data, and even in the posttest data. Here are some patterns that might be looked for in the intercorrelations among a measure such as English vocabulary. The "intercept" is a value extrapolated from the rates, to a point of no temporal erosion at all. It is a kind of reliability.

If there are systematic trends toward higher and higher one-year test-retest correlations, as there may be in some longitudinal studies, this may be interpreted as either a case of increasing intercepts with constant rate, which we currently favor (16), shown in Table 2, or as

Table 1

Cross-temporal Correlations of Equal Erosion Rate (.80) and Intercept (.90)

	GRADES						
	6	7	8	9	10	11	12
6	(.90)						
7	.72	(.90)					
8	.58	.72	(.90)				
9	.46	.58	.72	(.90)			
10	.37	.46	.58	.72	(.90)		
11	.30	.37	.46	.58	.72	(.90)	
12	.24	.30	.37	.46	.58	.72	(.90)

Table 2

Cross-temporal Correlations with Constant Erosion Rate (.80) and Increasing Intercepts

	GRADES						
	6	7	8	9	10	11	12
6	(.65)						
7	.54	(.70)					
8	.56	.58	(.75)				
9	.58	.60	.62	(.80)			
10	.59	.62	.64	.66	(.85)		
11	.61	.64	.66	.68	.70	(.90)	
12	.63	.65	.68	.70	.72	.74	(.95)

Table 3

Cross-temporal Correlations with Constant Intercept (.80) and Increasing Rates. (Indistinguishable, without Information on Reliability, from Table 2)

	GRADES						
	6	7	8	9	10	11	12
6	(.80)						
7	.54	(.80)					
8	.56	.58	(.80)				
9	.58	.60	.62	(.80)			
10	.59	.62	.64	.66	(.80)		
11	.61	.64	.66	.68	.70	(.80)	
12	.63	.65	.68	.70	.72	.74	(.80)
Rate	(.65)	(.70)	(.75)	(.80)	(.85)	(.90)	(.95)

a constant origin with increasing rate, shown in Table 3, containing identical values as Table 2 except for the diagonal. Insofar as the intercept conceptually corresponds to a synchronous test-retest correlation without memory for specific items, and is therefore like an internal consistency reliability, such reliabilities if computed on the same *S*s would be relevant to choosing a model. (In the STEP-SCAT longitudinal data, no incremental pattern seems indicated; Table 1 could be assumed, with some unevenness of reliabilities and intercepts for the yearly testings but of no orderly pattern.)

Shifts in schools, as between junior high and high school, may create greater erosion than the normal one-year erosion rate. Such outcomes as Table 4 should be looked for.

REMEDIAL OR COMPENSATORY PROGRAMS

In the previous illustration, the selection bias and the treatment effect operated in the same direction. In many remedial or compensatory cases the reverse is the case, and the effects of treatment and temporal erosion may be in the same direction. This probably means that unequivocal evidence of effects is rarer, but the analysis should still prove relevant.

One such case comes from the current ETS preschool longitudinal

Table 4

Cross-temporal Correlations with Junior High—High Break between Ninth and Tenth Grade

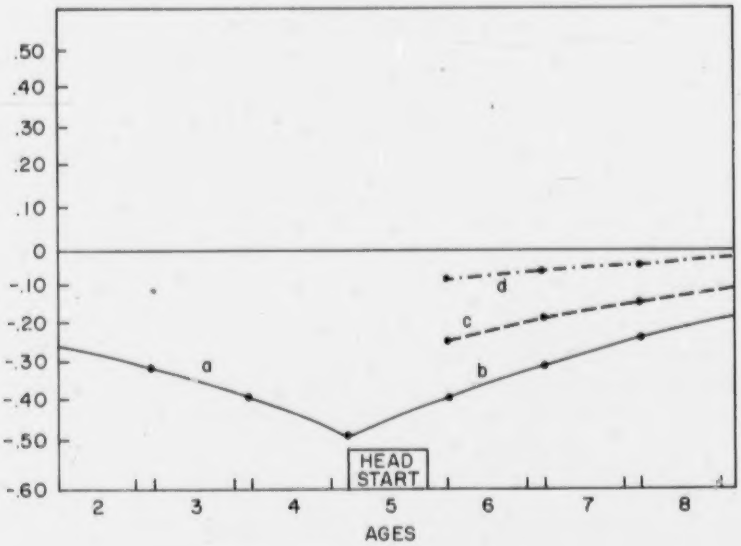
			GRADES				
	6	7	8	9	10	11	12
6	(.90)						
7	.72	(.90)					
8	.58	.72	(.90)				
9	.46	.58	.72	(.90)			
10	.30	.37	.46	.58	(.90)		
11	.24	.30	.37	.46	.72	(.90)	
12	.19	.24	.30	.37	.58	.72	(.90)

study in which some children receive Head Start. If Head Start is given to those who, on the average, need it most, as a compensatory program should be, then the pretest correlations with Head Start exposure are negative. A successful treatment makes this correlation less negative. Temporal erosion makes it less negative. Figure 5 plots such a situation. The values for lines *a* and *b* are those of Figure 3, except negative (.50, .40, .32, .256, .2048). Line *d* starts with a .30 increment, as for *d* of Figure 3, and this treatment effect dissipates at .80 (producing the reduced increments of .24, .192, .064, .0512). The net effect is for lines *d* and *b* to come closer together while both approach zero. If the treatment effect were to dissipate more rapidly, the net effect could actually be an increase in the negative magnitude of the correlation.

Figure 5 has been plotted with as sharp a peak as Figure 3. No doubt this presents an exaggerated view of the erosion and peak location problem. Probably the sharpest peak will come from selection decisions based upon individual pupil attributes. If the decision is based upon neighborhood or school attributes, the cross-temporal neighborhood correlations will be higher than person correlations and will show less erosion. The decisions are not apt to be time-specific as far as individual children are concerned. Longitudinal data give us the power to ascertain these facts.

Figure 5

Biserial correlation of annual September vocabulary tests with Head Start experience. Line b *is a clear-cut case of no-effect, line* d *a clear-cut case of effect.*



Problems with Data Limited to One Pretest and One Posttest

Imagine in Figures 3 and 5 that only one pretest measure and one posttest measure are available. In Figure 3, with outcome *b* one would not be tempted to claim a positive effect, whereas, in Figure 5, outcome *b*, one might be. The fact that treatment counters attenuation has made an outcome like *d* unequivocally an effect in Figure 3, but interpretable as rapid erosion in Figure 5.

Thus, for those instances in which the initial correlation and the treatment effect are in the same direction, treatment and attenuation have opposite effects, and a simple one pretest, one posttest analysis (5, 27, 2) is interpretable, albeit with excess conservatism. In the other instances, the one pretest, one posttest design is extremely vulnerable to mistaking erosion as a treatment effect, and the need for longitudinal data is extremely great.

Artifactual Sources for an increment in Treatment

The "plausible rival hypotheses" approach to quasi-experimental design demands that we look for likely sources of a correlation increment, as in Figure 3, other than a treatment. In Campbell and Clayton (5), it was argued that the co-occurrence on the same interview of the posttest and the ascertainment of exposure would create a higher posttest exposure correlation than pretest exposure correlation whether or not the treatment had an effect. In that case, the treatment was seeing an anti-antisemitism movie and the dependent variable was an antisemitism scale. In all panel studies, some persons are misidentified, different persons providing the pretest data than the posttest data. This lowers the exposure-pretest correlation but not the exposure-posttest correlation where exposure is retrospectively ascertained in the posttest interview. Furthermore, forgetting one has seen the movie, or erroneously reporting that one has, are attitude symptoms, and attitude measures occurring in the same instrument and testing situation always correlate higher than when the same measures are separated in time.

The same problem could occur in causal analysis in longitudinal studies. Consider another ETS interest, the impact of the *Sesame Street* children's educational TV serial. Here the longitudinal data of the Head Start study could be used, ascertaining which children have seen the series. The occasion of ascertainment should be kept separate from the testing program.

Correlational Analysis Where the Treatment Occurs in Degrees

In the *Sesame Street* and Head Start examples, and many others, one will have wide ranges in degree of treatment, number of days attended, or programs seen. There is no reason why the correlational analysis here described should not be employed, using the treatment as a continuous variable (with a mode, unfortunately, at zero). But for this analysis, the presumed correlation peak in the no-cause condition should be conservatively placed in the middle of the treatment period, as indicated in the discussion of decision times above.

Partial and Multiple Correlation, Matching, and Covariance Analyses

These techniques represent pathetic efforts to artifically reconstruct a zero pretest-treatment correlation by "controlling for," "covarying," or "partialling out" the pretest correlation from the posttest. As many have demonstrated (28, 18, 19), and the others have reviewed (21, 1, 6), these statistical procedures are inappropriate to the task. If the reader doubts this, let him apply his favorite analysis to the no-cause conditions illustrated in Figure 3 and Table 1. Erosion is not at issue here. Even if there was no cross-temporal erosion, there would still result non-zero pseudo-effects in our illustrated no-cause conditions, a significant positive increment or positive partial correlation in the Figure 3, Table 1 case.

REFERENCES

1. Brewer, M. B., Crano, W. D., and Campbell, D. T. Testing a single-factor model as an alternative to the misuse of partial correlations in hypothesis-testing research. *Sociometry*, 1970, *33*, No. 1, 1–11.

2. Campbell, D. T. The effects of college on students: proposing a quasi-experimental approach. Duplicated Research Report, Northwestern University, August 1967, 9 pp.

3. Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

4. Campbell, D. T. Reforms as experiments. *American Psychologist*, 1969, *24*, No. 4, 409–429.

5. Campbell, D. T., and Clayton, K. N. Avoiding regression effects in panel studies of communication impact. *Studies in public communication.* Department of Sociology, University of Chicago, 1961, No. 3, 99–118. (Reprinted in *Reprint series in the social sciences*, S-353. Indianapolis: Bobbs-Merrill, 1964.)

6. Campbell, D. T., and Erlebacher, Albert. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: a national debate. Vol. III* of *The disadvantaged child.* New York: Brunner/Mazel, 1970, 185–210. (Reply to the replies, 221–225.)

7. Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963, 171–246.

8. Cicirelli, V. G. The relevance of the regression artifact problem in the Westinghouse-Ohio evaluation of Head Start: A reply to Campbell and Erlebacher. In J. Hellmuth (Ed.), *Compensatory education: a national debate. Vol. III* of *The disadvantaged child.* New York: Brunner/Mazel, 1970, 211–215.

9. Cronbach, J., and Furby, L. How we should measure "change"—or should we? *Psychological Bulletin*, 1970, *74*, No. 1, 68–80.

10. Guttman, L. A generalized simplex for factor analysis. *Psychometrika*, 1955, *20*, 173–192.

11. Hilton, T. L. Intercorrelations among SCAT and STEP administered in grades 5, 7, 9, and 11. Personal communication, October 20, 1969.

12. Hooker, R. W. Correlation of the marriage-rate with trade. *Journal of the Royal Statistical Society*, 1901, *64*, 485–492.

13. Humphreys, L. G. Investigations of the simplex. *Psychometrika*, 1960, *25*, 313–323.

14. Janoušek, J. Comments on Campbell's "Reforms as experiments." *American Psychologist*, 1970, *25*, No. 2, 191–193.

15. Kendall, M. G., and Stuart, A. *The advanced theory of statistics. Vol. III.* New York: Hafner Publishing Company, 1968.

16. Kenny, D. A. A model for temporal erosion and common factor effects

in cross-lagged panel correlation. M.A. thesis, Northwestern University, 1970a.

17. Kenny, D. A. Testing a model of dynamic causation. Presented to the Conference on Structural Equations, University of Wisconsin, November 12–16, 1970b.

18. Lord, F. M. Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 1960, *55*, 307–321.

19. Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, *68*, 304–305.

20. Lubin, A. Time series analysis of repeated measures. Colloquium at Northwestern University, October 1969.

21. Meehl, P. E. Nuisance variables and the *ex post facto* design. In M. Radner and S. Winokur (Eds.), *Analyses of theories and methods of physics and psychology. Vol. IV Minnesota studies in the philosophy of science*. Minneapolis: University of Minnesota Press, 1970, 373–402.

22. Pelz, D. C., and Andrews, F. M. Detecting causal priorities in panel study data. *American Sociological Review*, 1964, *29*, 836–848.

23. Porter, A. C. *The effects of using fallible variables in the analysis of covariance*. Ph.D. dissertation, University of Wisconsin, June, 1967. (University Microfilms, Ann Arbor, Michigan, 1968).

24. Rickard, S. The assumption of causal analyses for incomplete causal sets of two multilevel variables. *Multivariate Behavioral Research*, 1971, (in press).

25. Rozelle, R. M., and Campbell, D. T. More plausible rival hypotheses in the cross-lagged panel correlation technique. *Psychological Bulletin*, 1969, *71*, 74–80.

26. Schmookler, J. *Invention and economic growth*. Cambridge: Harvard University Press, 1966.

27. Seaver, L. B. College impact on student personality as reflected by increases in treatment-outcome correlation. M.A. thesis, Northwestern University, 1970.

28. Thorndike, R. L. Regression fallacies in the matched groups experiment. *Psychometrika*, 1942, 7, 85–102.

Discussion

JOHN W. TUKEY
Princeton University

It is hard to have a discussion when one only has good ideas to discuss. As a charter member of the Society for the Suppression of the Correlation Coefficient, which used to have a little sign by which its members might know one another, let me begin by saying that I am prepared to applaud that part of Professor Campbell's discussion which said it was better to look at *t* or F or some other test of significance. As to whether it is better to look at *t* or F than to look at something with regard to the differences in means, I think he and I might have a side discussion some time. I do think that it is going to be important, as we move to use this new technique more and more, that we find out which way of measuring things gives us the best knowledge of behavior—gives us the best opportunity of projecting what treatments would be likely to produce.

I am not at all sure that the answer to this is the correlation coefficient. I am not at all sure that I can name any one thing that I think would be a four-to-three bet to be it. But it seems to me I would certainly have to look at Fisher's Z, as well as Pearson's r, even if I am tied to the correlation coefficient. The decay of Z might behave better than the decay of R. We don't know until we look. There must be lots of data, with no treatment variables, through which this sort of question can be looked at. I think the question of whether you should look at the covariance instead of the correlation coefficient is also up for grabs. Whether dividing by the other variance is a good thing or not is not as clear to me in this situation as it would be if all I wanted to do was to increase the stability of a measure whose meaning I didn't care about.

Whether we need to believe that things really follow a Gaussian distribution is always debatable. If we want measures of this sort, I think we need to know a little bit more about which parts of these distributions we are most interested in, particularly with the compensatory treatments. This might lead us to look at other measures. I could say a word or two of a more technical nature on this, but I think it would be wise for me not to.

Let me turn to adjustment and matching for a moment. I think the objections you have heard are in many cases well taken, but I think we also ought to bear in mind that adjustment may be okay if you don't believe it too much. The difficulty lies in believing that after you have done it you have adjusted things and the job is over.

It is not clear to me, however, whether there might not be circumstances where I would want to use adjustment and matching and then follow with Campbell's technique applied to the adjusted values—as in the famous discussion between Student and Fisher and the interjections by Sir Harold Jeffreys, it may not be a bad thing to use all the allowed principles of witchcraft and not just one set.

Problems that need to be dealt with in other ways arise when one adjusts to broad groupings. For example, one has the feeling that, having adjusted for some broad grouping, the adjustment task is over. This, too, is wrong, but fixable by other techniques when all that is the matter is the use of broad groupings.

I began to wonder a little during the presentation what the connection would be between what we are thinking of here and what is known as "superstandardization." At the moment, I think the only place you can find any discussion of superstandardization is in the report on the national halothane study.

Only some minuscule part of this audience probably will have had anything to do with things like standardized death rates, but there are techniques for answering the question: If you know the age composition or some other composition of two groups, how do you at least adjust the death rates to allow for this compositional difference? And, again, it is clear that in most cases it is better to adjust than not to adjust, and it is clear in some cases that it is wrong to feel that adjustment has settled everything.

In the national halothane study some 34 hospitals were involved. Halothane, if you don't know, is one of the most used anesthetics for surgical operations. There were adjustments for death rates by the hospital for various things, including age of patients, the severity of operation, and so on. It was very interesting to find that if you plotted the logarithms of the adjusted death rates against additive adjustment that you had already used for sensible reasons, the remaining regression was quite substantial. And it seemed to make arguable sense that you should make a further adjustment. In the original example, that amounted essentially to multiplying the first regression adjustment—the first adjustment which wasn't found as a regression adjustment—by about 1.6. You can argue quite awhile which of these sets of answers you think is more appropriate—which of the various things that might have influenced this situation were and were not taken into account in the first adjustment and might or might not be picked up by the second. But when you start comparing neighborhoods and things of that sort, I am not sure but what there might be some way to combine some of the ideas of superstandardization with some of the techniques we have heard this morning, yet I don't see how we can possibly avoid putting Campbell's technique to very serious use and testing it by seeing whether it does in fact show the things that are obvious to the naked eye.

Luncheon Address

Higher Education: For Whom? At Whose Cost?

CARL KAYSEN
Institute for Advanced Study

Currently, nearly 8 million students are enrolled in what we term "institutions of higher education"—the 2,200 or so nonprofit universities, colleges, and junior colleges that offer academic professional and semiprofessional training to high school graduates—and the figure is expected to press 10 million before the end of the decade. Total enrollment in these institutions has more than doubled in the last decade, a rate of increase much higher than the 25 percent growth of the decade before. It is even higher than the 80 percent growth of the 1939-49 decade, which covered both the radical change from prewar depression to postwar prosperity and the enormous surge of enrollments supported by the GI Bill of Rights.

Part of this growth reflects population growth and the related change in age distribution, of course, but the more significant part has been the continued increase in the share of each age cohort that finishes school and goes on to post-secondary education. Both the proportion of each cohort finishing high school and the share of high school graduates entering college has been rising steadily for nearly four decades, and the fraction of an age cohort entering college is now over 30 percent. A somewhat broader measure—the proportion of the population aged 18 to 21 who are enrolled as undergraduates in college—is available for nearly a century, and it shows continuous though varying growth. Its current level is somewhat over 40 percent, compared to some 31 percent a decade ago, 27 percent two decades ago, and just below 15 percent at the outbreak of World War II.

Meanwhile, graduate enrollment has been growing even faster than undergraduate enrollment. By the end of the decade graduate enrollments are expected to reach at least 2.5 million—the size of *under-*

graduate enrollments in the early 1930s.

The full economic costs of higher education are difficult to measure, and for both conceptual and statistical reasons. However, the specific outlays of institutions of higher education can be measured with some precision. These are currently on the order of $.19 billion per year. These outlays are financed about half from private sources and half from governments, with the federal government providing somewhat less, and state and local governments somewhat more than equal parts of the governmental share. About three-quarters of the private share comes from direct payments by students and their families in the form of tuition, fees, and room and board bills, with the balance from endowment income and endowments themselves, gifts, and grants.

If, as some economists would argue, a measure of the full economic cost of higher education should also include the cost represented by the foregone or "lost" earnings of the students, the bill might rise by another $10 to $20 billions, with the private share increasing correspondingly.

Figures in billions may numb the mind, I suppose, but what is in store for us is suggested by many other indices. It is interesting, for example, to compare the growth path of college enrollments as a proportion of the 18–21 age group with that for the corresponding proportion for high school enrollments in relation to the 14–17 age group. Roughly 30 years in time separated the two curves horizontally over much of the period since the beginning of the century. If this relation is maintained in the future, college enrollments will approach "saturation" about 1995.

There are other and less speculative indicators of things to come. Even now, states that are more prosperous and have extensive systems of public higher education show much higher proportions of high school graduates entering college. California leads and other rich states follow. However, the roughly 30-year gap between the times at which relative enrollments in high schools and colleges have reached the same level suggests a simple explanation for this kind of growth. The high school graduates of one generation want their children to be the college graduates—or at least the college students—of the next generation. Though crude, this account contains at least the germ of the truth. A more elaborate explanation would involve at least four factors, two private and two public.

On the private side, the first point one might make is that higher education, by and large, is a luxury good. This is not only to say that it is expensive, but that the higher the level of income, the larger the share of income it tends to claim. This proposition holds in respect to individual households both at a moment of time and historically, and also consistently enough so that it appears to hold, both comparatively and historically, for nations as well. Why this should be so raises a complicated set of questions beyond the analytical reach of the economist. A second point is more relevant: Higher education is the ticket of admission to higher levels of occupation, especially as measured in terms of status and income. As I have said elsewhere, "Some kind of advanced education, general or specialist, is increasingly a prerequisite to membership, not just in a small elite, but in the wide middle class of an advanced industrial society. We may say that in the United States today, and increasingly in the future, the public served by this aspect of the process of higher education is the whole middle class of our society. Some higher education is already a nearly indispensable ticket of entrance to middle class status for boys of working class origins. It will soon become only somewhat less indispensable to the maintenance of that status for those who were born in it" (1).

These two factors, mainly private, account for the steadily growing demand for higher education in a competitive, mobile, and steadily wealthier society. But higher education is not supplied via the private market, wherein enterprises are expected to meet any demand that is sufficiently profitable—which in practice means any large and growing demand. Higher education is supplied for the most part through the agency of governments, whose responses are determined by other forces. The demand for more higher education, for example, is increasingly expressed by the most effective elements of our society politically, and it is justified in terms of two highly prized and widely shared values: equality of opportunity and economic growth. It is this reinforcement of private demand by public justification that lends so much force to the drive for the further and even quicker expansion of the scale and scope of higher education.

Wide access to education is a major element accounting for social mobility. But though the college intake is wide and increasing, it still falls short in terms of equality of opportunity. As this audience knows, there is a serious discrepancy in the extent to which equally able high school students in different social strata go to college. Nearly 8 out of 10 students in the upper 20 percent of the ability distribution who finished school in 1960 entered college in the next five years, but the figure varied from 95 percent of those in the highest quarter of the income-status distribution down to 50 percent for those in the lowest quarter. At the other end of the ability distribution, an average of only 20 percent entered college. But this figure ranged from 50 percent for those in the top quartile of the status distribution to 15 percent for those in the bottom. Over all ability groups, 8 out of 10 of those from the top rungs of the status ladder entered college, compared with not quite 1 in 4 of those from the bottom. Similar, though not quite so sharp, discrepancies appeared in the extent to which those who entered college went on to receive the bachelor's degree four years later, with the figures varying from nearly 80 percent for those in the top of both distributions, to only 30 percent for the most able from the bottom quartile in terms of the social scale; and 66 percent for the least able at the top of the social scale, down to less than 30 percent for those at the bottom class of both distributions.

The most powerful justification—both practically and ideally—for continued and even accelerated growth in the size of our higher educational establishment is that such growth is the best way to diminish these discrepancies. It is, indeed, probably the only way. We have done much in many areas—if not enough—in redistributing plenty; the redistribution of scarcity is a grim task for a democratic society.

Moreover, expenditure on more higher education is not only a necessary cost for a more equitable society; it can also be seen as an investment in further economic growth. Sophisticated economic analysis generally finds that within the national perspective an increasingly better educated and trained labor force is a major input to the sustained increase over time of production per unit of input resources that our economy has enjoyed. The effect of such an expenditure can also be seen within the more limited perspective of governors and state legislators, who see in the growth in higher education a stimulus for local industrial development and the consequent growth in the population and power of their own sovereignties. The first of these views is not necessarily wholly correct, nor is the second wholly erroneous, and the second view is probably the more influential in the short run.

However, as the fiscal burden of sustaining the further growth of higher education shifts from the states to the federal government—with its more sophisticated bureaucracy, more intensive political struggle, and more continuous and wider public involvement—the extent to which the first of these arguments is valid becomes more important.

Two years ago, the Carnegie Commission on Higher Education sketched a program for sustaining the general growth of higher education and accelerating the opening of the system to those now heavily disadvantaged. Over the decade ending in 1976–77, the projected growth—along with a continued rise in costs—was seen as requiring an increase in total annual expenditures for higher education on the order of 2½ times, from the 1967 figure of 17 billion to 41 billion dollars by 1977. The contribution of the federal government was expected to rise almost four times, from $3.5 billion, in 1967–68 to $13 billion in 1976–77, and its share of the total to rise from a little more than a fifth to nearly a third. Moreover, updating these figures suggests a 1980–81 total figure of some $50 billions.

Before Congress can be persuaded to appropriate sums of this magnitude, two kinds of questions must be examined with great care. Are the social justifications for such expenditures from tax revenue sufficiently clear and strong to warrant the sums involved? Are there more efficient ways of achieving the same ends?

Though this is more a matter of speculation than analysis, it appears to me doubtful that the federal responsibility for increasing equality of opportunity in this particular direction will alone prove powerful enough to guarantee expenditures on this scale. There is no obvious "right" or "natural" growth rate for the process of improving access to higher education, and thereby economic opportunities, for the children of those at the bottom of the social scale. Competing and growing claims for federal expenditures on other aspects of this equalizing process—in primary and secondary education, in welfare, in medical care—will limit the weight given to the single area of higher education. To the degree that all such programs involve income redistribution via the federal tax system, the high correlation between the distribution of income and political power will limit their total extent. Thus, the argument for investing in higher education as an inducement to general economic growth becomes critically important. If a convincing demonstration can be made that what is at stake is not merely the redistribution of income and opportunity, but a *collective* investment that will yield higher incomes for all, the whole question could be viewed in a different, more favorable context.

It is just on this point, however, that considerable scepticism is in order. The by-now standard economic analysis of education as an investment involves two steps. First, the lifetime earnings of those with more education are compared with the earnings of those with less—allowance being made for the cost of education, including the earnings foregone during the period of education, the differing time patterns of income in different occupations, and the like. Such comparisons provide a basis for calculating rates of return on investment in education.

But these, of course, are private rates of return to the individuals who receive the education. And the question of whether the aggregate of such private returns constitutes an appropriate measure of social return remains. One way this question has been answered is by correlating aggregate productivity per unit of input with the aggregate "stock" of education embodied in the labor force, both over time within individual countries and between countries. But interesting as such analyses are, they depend on assumptions about the mechanisms connecting individual with social returns, which is the very point in question. Another way has been to take an arbitrary fraction—for example, ⅔ or ½; but this is also question begging.

The nature of the problem can be put in terms of two sharply contrasting and simplified descriptions of what, in economic terms, higher education seems to do to contribute to productivity. In the first projection of this type, we conceive that each educated person receives—inter alia—specific training of some sort, whether in a technical skill, an intellectual process, or some general learning ability, which makes him a better worker in a specific occupation or range of occupations. Further, we assume these skills cannot generally be gained in other ways, for example, by experience on the job. This gives us a naive picture that might be labeled "education as training." According to this model, the aggregate of individual returns net of costs would indeed be a measure of social return.

The contrasting picture would start with the fact that, at any one time, the number of jobs with high pay and high productivity is limited, and some selection system must allocate access to them. In less mobile and dynamic societies that function is performed to a great extent by the kinship system. However, in our more mobile society, with its changing occupational structure, to a large degree it has become the task of the system of higher education—passage through which is to provide a certificate of admission to the higher levels of the occupational structure.

According to this view, then, higher education functions as a selection system: Both admission procedures and the courses of study, even though they may be devoid of specific content relevant to occupational performance, are seen as representing an obstacle course. And those who make their way through it are seen, in contrast to those who do not, as possessing the qualities of intelligence, energy, application, and persistence that are indeed necessary to effectively perform the tasks for which graduates are certified. If this description correctly characterized the functioning of higher education, the question would then be: Is such an elaborate, expensive, and time-consuming process needed, or can the tasks of selection and certification be performed in a simpler and easier way? According to this model the social return for higher education—apart from the satisfaction it provides to those who receive it—would be very little, and would be limited to its value as a selection system.

In practice, the system of higher education contains elements of both of these models, and its actual functioning is a mixture of certification, selection, and substantively useful training. Different institutions and programs will vary in the extent to which these elements are operative. However, observable trends in the educational system itself, and in its relation to the job market, point to a rise in the component of selection and certification in relation to substantive, occupationally relevant training. As an increasing proportion of each age cohort enters college, and as a much lower but also increasing proportion graduates, the tendency to upgrade the educational requirements for entry into the higher occupations rises. Technical jobs become semiprofessions, and even professions, for which graduate as well as college training is required.

From the point of view of the hiring employer, this upgrading is a costless process, so long as the proportion of graduates keeps increasing. A comparison of the occupational distribution of college graduates in the United States with those in a country like Sweden—which also has very high per capita incomes and growth of productivity but comparatively small fractions of age cohorts entering post-secondary education—makes this point strikingly, by showing how much further down the occupational ladder our college trained are spread. This leaves us with a difficult evaluative task. We need a critical quantitative estimate of the extent to which increasing the rate of growth of higher education would indeed be an efficient way to increase the output of economically useful skills, as against simply providing a larger supply of college men and women to be absorbed by a corresponding upgrading of job standards—ignoring, of course, the noneconomic benefits of the increased education and its simple enjoyment as a consumer good by those who experience it. Research that would contribute to such an estimate is only just beginning and more is needed. But the simple case for an increase in social benefit proportionate to the increase in the scale of higher education is hard to accept on the basis of present evidence.

There is also much to be said about the consequences of continued steady growth in the size and scope of higher education from the narrower viewpoint of academia itself, or at least that part of it to which we who think of ourselves as the custodians of its inner mysteries belong. It is obvious that the rapid growth of the recent past has added to our incomes and our prestige, although this has been as much the effect of the growth in research as in education proper. Growth in our numbers, and even more in the numbers of our students, is making us, at least potentially, a political force of some consequence. Whether these changes are blessings I leave to you to decide.

Within this "core" to the system, we have come to rely on higher education to perform a variety of functions. They can be categorized in four classes:

1. The transmission of knowledge to the new generation, including
   (a) "general culture," and
   (b) technical training for a variety of professions
2. The creation of new knowledge and its integration into the present body of knowledge; that is, research and scholarship (It is, of course, the graduate programs leading to the Ph.D. that are typically the source of the technical training for this second function.)
3. The application of special knowledge to the solution of social problems in the larger society
4. The socialization of the young

This last function is clearly connected with the first function, especially the first part of it, but it is less a matter of knowledge and more one of sentiments, attitudes, values, and the formation of enduring personal associations.

As for certification, it could be recognized as a fifth function, or, alternatively, seen as a by-product of the first and fourth.

The university, the central institution of the system, typically performs all of these functions. Other institutions—the liberal arts college, the two-year junior or community college, the independent technical or professional school—perform varying combinations of them, but typically none makes any substantial contribution to the second function, and few to the third. Only the university trains those who staff the university, and now trains most of those who man the whole system. Further, this function is highly concentrated in the two to three dozen universities that produce most of the Ph.D.s and an even higher share of the serious scholarly and scientific work. These central institutions have grown much more slowly than the system as a whole, and, indeed, as far as the growth of the last two decades in undergraduate programs goes, the bulk of it has occurred in other places. This will be even more true of further growth in the future.

The institutions in which the great growth has occurred—the state colleges, the new branches of the state universities, the junior colleges, the municipal colleges and universities—are primarily engaged in teaching, socializing, and certifying; the rest is for the most part outside their scope. But the bulk of their faculties are still produced in the universities, absorb the culture of the universities, and operate with purposes and in terms of models that will often diverge widely from those of the student bodies in the institutions where they teach. Consciously or unconsciously, they seek the scholar in the student, and they think their highest task is to find those who can go on to the next higher stage of the educational ladder.

The effect of the central model of the university on the whole system has been to subject every institution to pressures to become a university and every teacher to become a "researcher." Thus, demand for expansion in the lower parts of the system has induced proportional growth in the whole, whether or not this growth is necessary or desirable. Good research and scholarship are not as readily multiplied as are programs of technical and professional training, and the attempt to do so raises the costs of the whole process and diverts resources and attention from the goal initially sought.

Further, as the system undergoes growth, demand presses on capacity. In consequence, scarce places at all levels tend to be reserved for those coming immediately from the next lower rung of the ladder, and "dropouts" of all kinds find reentry into the system difficult. The system thus favors those who, so to speak, go through it without pause from kindergarten to the Ph.D.

Finally, the current process is wasteful in that attrition rates are high at all levels. The best available figures show that nearly half the students who enter college do not earn a bachelor's degree; more than 1 in 6 of those who try for a master's or first professional degree fail to achieve it; so do half of those who try for a doctor's degree. Surely a system as wasteful as this should not be expanded at high costs and public expense without a search for alternative ways to achieve the same purposes.

One alternative to the present path of development in higher education is simply to halt, or at least drastically slow down, public financial support for its further growth. Aside from the undesirable consequences for existing institutions and programs such a course would bring, this simply is not a politically live option if the analysis of the forces making for growth offered above has merit.

Another more helpful alternative would be to seek to separate the functions of socialization and general certification, and to provide educational programs appropriate to this separation, including the development and staffing of institutions suited to perform these two functions. If this could be done, both the volume and the allocation of the other tasks of the system could be examined in broader terms than is now possible.

The development of junior colleges may be viewed as an effort to follow just this path. But while they have grown rapidly in number and enrollment, they lack an essential ingredient. As they do not offer a "college degree," they cannot satisfactorily perform the certification function. Some new scheme is needed.

I propose that a set of three-year programs, leading to the degrees of bachelor of arts or bachelor of science, be developed as the standard college program, which all high school graduates who could satisfy some relatively broad admission standard would expect to pursue. The programs could have varying combinations of general education and technical training for a variety of professions and semiprofessions. In principle, these same programs would be open to anyone who could meet the admission standard, whether or not he was a high school graduate, and whatever the recency of his secondary education. This kind of approach would become the basic program for the present state colleges, "branches" of state universities, and municipal colleges; and these institutions would be staffed chiefly by a faculty who were considered, and considered themselves to be, primarily teachers, not scholars. As is now the case in community colleges, at least some portion of the teaching would be done on a part-time basis by professional practitioners engaged chiefly in other occupations, whether as economists, chemists, or city planners. As new institutions of this type were created, they would follow the locational pattern of the state colleges, and their branches, and so forth, that permits a major part of the student body to live at home (or elsewhere close by) and to come to school as commuters. These institutions would perform their functions best if they were organized and staffed independently of the higher levels of the system.

Further, I propose something very much in keeping with a proposal by Dr. Machlup earlier in this conference: that the standard high school course be reduced to three years. This would in itself redefine the school dropout problem in a significant way, as well as provide welcome relief in many crowded urban schools. In combination, these changes would define a program of more or less universal college education that would occupy the same time span as high school and lower division or junior colleges do now. This in turn would provide a saving of the order of 20 percent of the present costs of higher education.

The community college, with an open admission policy and a wide variety of programs available to either full-time or part-time students would still be needed. It would serve as an institution for adult education for both vocational and cultural purposes, and also as a means for some who had not previously done so to achieve the standard of admission for the basic three-year college. Moreover, it should be possible for many community colleges to share much of their facilities and faculty with the basic three-year colleges.

At the next higher level above the three-year college there should be a range of institutions providing technical and professional training over the whole spectrum, from agronomy and architecture to teaching and veterinary medicine, and including training for teachers in the basic colleges. The duration of these programs would range from one to three years. This should be the function of a substantial number of the present universities, as well as the better colleges, especially those that have already developed master's degree programs, or that could expand and change to perform these tasks. These institutions would also be involved in social problem solving, in terms, so to speak, of the clinical work of their professional schools. They should also provide refresher courses and continuing professional training for those in the work force. Finally, the central universities could concentrate their efforts more strongly on research and scholarship and on the training of faculty both for themselves and the larger number of professional training institutions. They would also continue to offer professional training over a wide spectrum.

In the task of training the next generation of faculty it would be important to recognize two very real differences in emphasis: those appropriate to the training of scientists and scholars, who should be judged by the value of their contributions to the stock of knowledge, and those appropriate to professional practitioners or teachers of difficult subjects, who should be judged by their effectiveness in applying and transmitting knowledge. One way of insuring such recognition is the creation of different degrees for the two.

With respect to the first level of post-high school education, this scheme has a double intent. It is, first, to lower even further the gradient between high school graduation and college, and to provide the first level of certification on relatively easy terms, as well as to make it more and more widely available. At the same time, the present gradient between college and what comes after in the way of professional training, including that of scientists and scholars, should be raised. The aim here would be to discourage the spread of higher level certification as an occupational prerequisite, and the consequent greater reliance on training on the job, including part-time training while working, refresher courses after employment, and the like. Such a move should, in turn, help break the temporal rigidity of the educational scheme and permit more interweaving between occupational experience and formal training, a pattern that was much more common a generation ago than it is now.

The scheme also embodies an explicit recognition of the increasing demand for vocationally oriented training, and the corresponding decline in the concern for "general culture." This, too, is a concomitant of the desire for universal higher education. We may not approve of it, but we must recognize that, to date, whatever efforts have been made to resist it have failed.

We may expect, also, that as basic college education is made available on a more or less fully subsidized basis to nearly everyone, the case for making tuition charges for the higher educational levels bear some relation to costs, will become progressively stronger, and the trend will be to provide loans available to all and repayable on the basis of future income—the so-called contingent loan fund plan—rather than scholarships. The interest rate embodied in the repayment scheme will then serve as a convenient vehicle for reflecting social judgments on the value of post-college education (in our new terms) to society over and above its value to the individual. In general, to the extent that such a loan scheme and its corresponding charges are used, a much greater reliance on something like market principles to determine the levels of activity at these stages of the educational process becomes possible. This in itself would provide a considerable step forward over present methods of resource allocation for education.

One purpose of all this is to differentiate, one from another, the ever-wider provision of what I have called basic college training and a corresponding growth in the whole system of higher education. Unless this differentiation is made, the costs of achieving college education for everybody will be enormous, and the waste in trying to do so equally great. A second purpose is to avoid a spread in the tendency toward formal certification for every skilled and high-status occupation. Unless this tendency is checked, we may defeat our end of broadening economic and social opportunity. In the absence of such changes, a system which has in the past been an important channel of intergenerational mobility may become a significant barrier to intragenerational mobility among occupations. And in a rapidly changing society, the importance of maintaining such mobility grows. Nor is this scheme really so radical as it might at first appear. As I have projected it, it serves mostly to underline some favorable current trends, and to warn against other trends I find threatening or unfavorable.

Changes in these directions do not require the wholesale imposition of a plan but could be achieved by appropriate incentives. And here, as in so many other current social concerns, it is federal money which can provide the incentive. On just what terms it is offered, and to whom, will for better or worse answer the questions that I posed in the title of this paper. But while the power of the federal purse can provide the energy for this or other reform schemes, the best thought of the academic community is needed to guide them.

REFERENCES

1. Kaysen, Carl. *The higher learning: the universities and the public*. Princeton, N. J.: Princeton University Press, 1969, 9–10.

General Session

Problems in Implementation

Social Accounting in Education: Reflections on Supply and Demand

DAVID K. COHEN
Harvard University

The Issues

The purposes of information systems in education are no different from the aims of social accounting in health or welfare. The systems are regarded as ways to make planning more rational and government more accountable by monitoring individual behavior and institutional performance. The underlying notion is that better information would improve the management of public institutions, make delivery of service more effective, render the production of benefits more efficient, and increase consumer power.

Given these similarities, it is no surprise to find that the political problems in social accounting are quite uniform. To judge from the last five years' debate, there are two chief issues: New information systems might further reduce the limits of privacy, and institutions might successfully resist the collection of data on their performance.

In education, more attention has been focused on the second problem. In part this has occurred simply by default. Children are accorded an almost entirely dependent status in the United States, and the ascription of such status naturally reduces concern about the protection of personal freedoms and civil rights. Because children are regarded as incomplete members of the polity, public institutions are permitted to probe their performance and regiment their behavior to a degree unthinkable in adult civilian society. The absence of much concern about the impact of educational information systems on the privacy of persons only reflects this attitude.

The other reason why most attention has been focused on the resistance to social accounting in education is that the resistance has been front-page news. Both the National Assessment of Education Progress and the *Equality of Educational Opportunity* survey (1) foreshadow social accounting in the schools, and both cases generated lively controversy. Local nonparticipation in the *Equality of Educational Opportunity* survey was massive, and there was rather a nasty struggle over the content and objectives of the National Assessment. Both controversies revealed an unmistakable resistance to scrutiny on the part of the public schools. This helped to solidify the impression that the major barrier to effective social accounting in education is getting the systems established.

This notion is consistent with most of the assumptions that underlie the movement for social indicators. Chief among these is the view that one of the principal obstacles to better institutional performance* is the absence of adequate planning, and of an adequate information base for such planning. While no one who has thought seriously about social accounts would minimize the barriers to their establishment, almost everyone seems to believe that if information were available it would be a major force for change. This, in turn, rests on the view that information on institutional performance has or could have an important influence on decisions.

No one could doubt that lack of information is an obstacle to change—but is it central? Is there any evidence that the schools would use the results that information systems spew forth? Do we suffer from a short supply of information or from a minimal demand for it?

I strongly suspect it is the latter. The deepest political problems in social accounting probably lie on the side of demand and consumption, not on the side of supply. On the schools' part, this arises from the fact that they are really not geared to utilize information on institutional performance. The organizations' incentives and structure rest upon other values. The schools' resistance to the Coleman survey or the National Assessment was only one symptom of their underlying inability and unwillingness to utilize such information.

But the matter reaches well beyond the schools to the general problems of information use in the political process. Most discussions of social accounting in education seem to assume that the output would serve both as political intelligence for the populace and management intelligence for the institutions. It certainly is true that the systems seek to improve management and "production" within government and to increase its political accountability. But in itself, would information accomplish either end?

---

*This may involve effectiveness, efficiency, or better management. In a paper this general, there is no need to distinguish among them.

Perhaps not. One might argue, for example, that the effect of these systems would be only to further clog the channels of political intelligence and weaken the links between school managers and their constituents. After all, while the revolutions in communication technology have vastly increased the amount of available information, there have been no comparable innovations in its social consumption, especially in public life. How is the deepening sea of information to be organized, interpreted, and brought to bear on decisions about the use of public resources? The established school interest groups have some capacity to utilize information, because of their organizational resources. But what of the citizenry, which is supposed to govern education? Will more information make them even more dependent on the existing institutions and further weaken their independent power as consumers, clients, and constituents? Will it not strengthen the power of managerial elites at the expense of democratic control? Will increasing the information flow further contribute to the growing sense of mystification, estrangement, and imperviousness which surrounds our institutions? Or to widening the disparity between the ability of affluent and poor people to cope with public institutions? These issues have not been probed.

My view, then, is that the two central political problems of social accounting in education are the dramatic absence of much institutional demand for the information and the lack of much consumer capacity to manage, control, or digest the products of social accounting. The most important issue is not how to establish new information systems, but how to assure that the systems' products would have some other purpose than the amusement and occupation of people like ourselves.

The remainder of this essay amplifies these ideas. First, I explore the relative importance of demand and supply. Second, I speculate on the consequences of creating major new sources of information supply when demand is minimal and consumers' utilization capacity is nearly absent. Third, I try to identify and evaluate the main alternatives that might increase demand and the capacity to utilize information.

In all this, several important political issues—or issues with political import—are either ignored or treated in passing. One is the question of what utilization of social intelligence might reasonably be expected from institutions and consumers in a large and diverse society. There is no more important issue than this, because these expectations are the basis for judgments that particular institutions work well or poorly. Although I have little doubt that in education they work poorly, much more thought will be required before we can talk sensibly about how much better they ought to become.

A second issue has to do with the technological viability of social accounting in education. What would be measured, and why? If the essential outlines of the learning and socialization processes were known—in economists' terms, the educational production function—this would be less difficult. But we do not know this, which leaves the awkward problem of deciding to measure things on the basis of either expert opinion or social consensus. There are many potential benefits, of course, in having recurring measures of status and change, even on those things we only *think* are important. But there also may be serious disadvantages. Suppose an information system turned up a considerable number of inequalities in some educational "input," and as a result much time, effort, and money was spent equalizing the differences. But is this worth it, if the inputs later were found to be unimportant? Or, to put the problem more broadly: What we measure in a national information system on schools will assume enormous importance, simply because it is being measured. Does it make sense to accord such political status to information whose real importance is dubious or unknown?

I raise these issues only to indicate that any full assessment of the political problems with social indicators in education should consider them. Unhappily, space constraints mean that I must pass over them for the time being, in order to attend to the more general issues of supply and demand.

The Absence of Demand

What would be required to show that I am incorrect, and that the main problem was supply, not demand? One important line of evidence would be repeated examples that the schools have employed available information to improve their performances, or created the necessary data. If such cases could be turned up, we would also be able to identify those elements in the public schools' organization that impel them to utilize information as a means of self-correction.

Nothing of the sort seems to be possible, however. To begin with, there is little evidence that the public schools utilize information on their own performances to improve operations. The most impressive example of this arises from contrasts between the history of the schools' "improvement" during the last four or five decades and the history of research on the effects of these improvements.

Ever since the turn of the century, the growth of American education has rested on the premise of some identity between the interests of the school professionals and students. The history of the last half-century in education might well be written in terms of shrinking class size, rising teacher qualifications, growing specialization within the educational professions, and increasing investments in public schools. The school professionals have pressed these changes with considerable success, and always with the belief that they would benefit students.

It is no surprise to discover that as these changes occurred, educational researchers sought to discern their impact. The result was a veritable avalanche of studies concerning the effects of such things as class size, teacher experience and qualification, school size, and educational expenditures on students' achievement. Yet, as J. M. Stephens pointed out in a recent review of these studies (2), the results were almost uniformly negative. Most of the changes which were supposed to make good schools from poor ones seemed not to make good students from bad ones. Class size, teacher experience, school expenditure, teacher qualification, and school size almost never affected students' achievement.

The accumulation of these studies seriously undermines the notion that the school professionals' interests are identical with childrens'. But this seems to have had not the slightest effect on school policy or practice. Indeed, despite the confirmation of these results on a grand scale by two massive national surveys within the last decade—Project TALENT (3) and the Coleman report—the education professions continue to assert that the only real barrier to improved education is the absence of adequate resources. The schools have either dismissed the results as bad research or behaved as though they did not exist.

It might be objected, however, that this example is unfair. Most of the research in question was unrelated to particular efforts at school improvement, was published in obscure journals by even more obscure researchers, and presented no alternative paths for action. On this view, a better example would center in the efficacy of schools' endeavors to monitor their own efforts to upgrade performance. Perhaps the outstanding case of this sort is the evaluation of programs to improve education for disadvantaged children, funded by Title I of the 1965 ESEA.

The results from most Title I project evaluations are even more discouraging. For one thing, they are in no way related to decisions about program design, planning, or funding. In almost every case, evaluation appears to be an entirely separate activity, the results of which are unrelated to the decision-making process. But even if they were, the quality of the evaluations is such that the feedback would have little effect. The overwhelming majority of evaluations simply are not designed to yield information either on gross program effects or on differential project effectiveness. They are mechanical, crude, and sterile; they are, in short, designed to satisfy a requirement for receiving funds, not to discover what best serves the interests of disadvantaged children.*

What is more, the results are not used in schools' relations with their clients and constituents. I have been able to find few instances in which evaluation results were made available to the populations at whom the programs were aimed. Indeed, there is by now a record of considerable resistance on the schools' part to releasing the results of evaluation, even to those established citizen advisory groups established by law or regulation under Title I. This, of course, is only one manifestation of a much broader pattern of behavior among local educational agencies: They are reluctant to make public much information about institutional performance.

This is not to say that the schools do not disseminate information. They do. Their initiatives in this connection, however, are ordinarily confined to those occasions on which public support for school programs must be organized to raise new monies. And even on such occasions, the information stays well within the bounds of those criteria enshrined in professional standards. The schools' "needs" and "successes" are related to the age of facilities, the qualifications of teachers, and so on. Other information, which might illuminate per-

---

*There are several reviews of evaluation practice in Title I programs. The most comprehensive is Wholey, J., and others, *Federal Evaluation Policy* (see references). I reviewed the issue in depth for one state, in "Public Education," in *The State and the Poor*, Beer and Barringer (Eds.) Cambridge, 1970, and generally in "Politics and research: The evaluation of social action programs in education," *Review of Educational Research*, April 1970.

formance differences among schools and school systems—such as test scores, track assignments, or post-school work or education—remains a mystery.*

The impression that emerges from all this is that public education agencies maintain a virtual monopoly over information on schooling. In addition, the available information is cast in terms that suit the interests of the educational professions. The schools themselves exhibit a deep antipathy to creating or utilizing information on institutional performance. What is more, they provide little information to clients or constituents, and none of it would challenge either the existing management of the enterprise or its definition of educational quality.

The evidence on these points could easily be multiplied, but there is little purpose in extending such a dismal tale. The important question is why the institutional demand for information is so low and consumers' capacity to manage it so underdeveloped.

The answer is that there are few if any incentives to utilize information on institutional performance within public education. One reason for this has to do with the character of the incentives and constraints within which education professionals work. The public schools are essentially a public employment system—a civil service. The criteria of personal advancement in such systems is defined largely in terms of standards created by the professionals involved, and very typically center in length of service and level of professional training. Thus, the focal points for competition among teachers within systems are almost exclusively bureaucratic—the amount of work toward advanced degrees, the extent of service in such nonobligatory tasks as curriculum committees, activity in professional organizations and activities, and sometimes specialization in a subject-matter area. The rewards include salary, promotion, and autonomy. None of this has anything to do with individual or institutional performance.

There is competition among schools and districts, but, as might be expected in a civil service employment system, this is not unlike that which occurs within school systems. The object for schools is to gain a larger complement of personnel whose attributes are desirable in terms of professional values—chiefly degrees and quality of school attended. The schools and systems that have more people with more such attributes are generally regarded as superior.

---

*I do not mean to suggest that such information is deliberately suppressed. Only a little is; most of it is never collected or analyzed.

Finally, status is attained not by making better students but by having them. The "better" schools and school systems are not those that take their students further from where they began, but those whose students go farther because they started with an advantage. This does not reflect any invidiousness peculiar to the educational professions—it simply mirrors the dominant social status system. What is more, there are few alternative upward routes within the school system. People who begin with low status and credentials cannot rise swiftly in public education—by becoming influential or wealthy—as they might in higher education, business, or crime. As a result, the main paths to advancement are either through serving time or gaining political power within one of the bureaucracies or professional organizations.

Thus, all the constraints on employment for school professionals are unrelated to individual or institutional performance. But this is hardly the only reason why information on institutional performance is neither sought nor utilized. Another important consideration is the existence of an ideology that identifies school performance problems with the clients, not the institutions. The schools operate on the explicit assumption that the sources of children's failure in school lie with the students, their families, and their social inheritance. Although this is not the place to explore the sources of this ideology, it is worth noting that it flowered as the cities' population was swelled by domestic and foreign immigrants.

The ideology is manifest in the extensive information system the schools do maintain. Although it provides no data on the performance of schools, there is an abundance of evidence on the performance of students. Pupils are tested for intelligence and achievement, graded on academic effort and standing, and rated on a bewildering variety of personal and character attributes. They and their parents are regularly apprised of these tests, grades, and ratings, and precautions are taken to make sure that the information is noted at home. All of this, of course, proceeds on the assumption that the source of children's academic difficulties lies outside the schools. The school information system contributes to this notion (indeed, it co-opts parents and children to it), as do the various "sciences" of education. Schools are not given report cards—they are not tested, they receive no grades, and their social, economic, or academic standing is never threatened for nonperformance.

This is not to say that there are no potential countervailing forces: At the local level, schools are politically accountable to the public; there are independent accrediting organizations; and state and federal agencies have some responsibility for insuring quality in local schools. This is to say, however, that these potential countervailing forces have little effect. Their impotence is the third major reason why public education neither demands nor creates information on institutional performance.

State education departments and the independent accreditation groups, for example, have established minimum quality standards for schools. They are backed up by sanctions, and when school systems fail to comply, more or less drastic penalties are invoked. But what are the standards? Do they involve institutional performance, or management? Upon inspection it turns out that minimal standards are defined almost exclusively in terms of the school professionals' criteria of quality—teacher experience and education, adequate facilities, class size, and so on. Moreover, the state agencies and the accreditation groups are staffed almost entirely by persons drawn from the school professions, who therefore share the commitment to professional standards. As a result, these institutions tend only to reinforce the assumption that the only relevant measure of institutional performance is implementation of professional standards.

Lay control at the local level also is constrained. The professionals who control the educational enterprise have developed a system of distinctions between policy and practice which keep laymen's hands pretty well out of the machinery. In addition, the sort of laymen who find their way to boards of education through some citywide selection process (elective or appointive), usually have enough other things on their minds to keep them from making serious trouble for the staff. And even if they didn't, the professionals serve as the sole staff for school boards, which assures that no countervailing power could emerge within the bureaucracy.

A final consideration is that the mechanisms for information use among the schools' clients and constituents are fairly primitive. The PTA is, for all intents and purposes, a captive of the professional associations at the state and national levels, and at the local level it operates in "partnership" with the school authorities. Further, the insulation of education from "politics" minimizes the constraints on schools exercised through the electoral process. School board candidates usually do not run with party identifications, and while they may build personal organizations after election, typically these are not large or strong enough to gather, process, and disseminate information that might undermine the schools' monopoly.

Within the structure of public education, then, there is neither countervailing power that might compel the schools to utilize information differently, nor sources of counterinformation that might challenge the schools' monopoly. Some potential checks exist, but the organizations have so completely assimilated professional standards that they have a precisely contrary purpose. Instead of promoting diverse standards of quality and competing information, they serve only to check deviations from the existing orthodoxy.

Is there any reason to believe that merely increasing the supply of information would change this situation?

The Consequences of Greater Supply

There are three reasons for an affirmative answer. One is that social accounting in education would produce information about technical improvements which would generate their own pressure for adoption. A second is that social accounting would reveal inequities in outcomes and the allocation of resources, and thereby multiply pressure for change. A third is that social accounting would become a countervailing information source, challenging the schools' monopoly in this area.

Of these, only the first point is clearly incorrect. Social accounting would be an unlikely source of information on technological innovations, since its purpose is to measure status and change on certain broad social indicators, not to identify particular innovations and evaluate their consequences.

It is more difficult to quarrel with the other two ideas. There are more than a few cases in which the presence of information has made a difference in government. Where would the *Brown* decision have been without the evidence on the effects of segregation? Or the reapportionment cases without the U. S. Census of Population? Or economic planning without data on productivity, prices, employment, and consumer behavior?

Similar examples can be produced for information as a source of countervailing power. Federal census information on population and housing has been turned to advantage by advocates of social legislation, outside and within the government. Information on civil rights compliance published by the U. S. Commission on Civil Rights has typically been at variance with other official information on the subject, and it has been useful to groups pressing for more vigorous enforcement of anti-discrimination laws. Labor and management see to it that the federal government collects and publishes information useful to their respective views of the economy.

But the common point in both sets of examples is that information alone would have little effect. It seems to become important when appropriated by existing political interests. Apart from muckraking (which produces horror stories of a sort unlikely to emerge from a system of social indicators), an estimate of the likely effects of information is really a judgment about the strength of contending forces in a given political arena.

There are few likely sources of such strength within state and local school systems. Are there other potential users of social accounting?

At the local level, the main hope seems to be those community groups and school reform agencies which have been struggling with the schools for the past 8 or 10 years. By all past standards these conflicts have generated absolutely unprecedented amounts of information on school problems, and considerable pressure for change. For the most part, however, this has been like water on a duck's back. The public schools have quite effectively ignored the information and resisted the pressure. The chief results have been a really remarkable series of changes in the ideological scenery (as reformers shifted their ideas about what should be done with the defeat of each earlier notion) and a substantial increase in the sense that public institutions are unresponsive. It is difficult to see how adding more information would change anything.

The other potential user of social accounting lies with the federal government. There is a body of thought which holds that America is ruled increasingly by trained managerial elites, not by the untutored, contending interests commonly discussed in textbooks on politics. As the management of public institutions falls increasingly to such technologists, the domain for "rational" decision making is thought to grow. Rational decisions, of course, require sound information.

For better or worse, this vision hardly squares with the facts, at least in education. There was a brief spurt of interest in scientific social planning in the mid-1960s, with the advent of PPB* systems in the Department of Health, Education and Welfare. There has, however, been little growth since then. The staff is too small to carry out the requisite analytic work, and the information base for it is mostly lacking. The data required for social planning activities on the PPBS model–program evaluation and comparative program effectiveness studies–simply do not exist (4). A system of social indicators could not provide them.

In my view, however, this is probably an incorrect way to view the likely consequences of a national system of social accounting in education. Although it would not lead to a radical change in the character of federal decision making, a well-conceived system could serve the same function as any census. It could identify existing inequalities in school outcomes and the distribution of resources and provide evidence on their trends. Such information might even have some impact on decisions about federally-sponsored school programs, and it might weaken somewhat the information monopoly currently enjoyed by the public schools.

Nonetheless, a system of social accounting would not have a major effect on the quality of state and local educational decisions or the performance of educational institutions. The federal share of public elementary and secondary school revenues is less than eight cents on the dollar, and federal influence on state and local decisions is comparably small. Even if a national system of social accounting were adopted, federal leverage is not sufficient to affect either the demand for or the capacity to utilize information among states and localities. Thus, while it is understandably attractive to focus on the relatively more flexible federal bureaucracy, the real problems lie elsewhere.

Stimuli to Greater Demand

I have identified two critical barriers to the use of social accounting at the state and local level. One is the absence of any incentives internal to school systems which would create a demand for the products of social accounting. The other is the absence of any countervailing forces, ones which might either use new information to affect school policy or use their influence to affect the schools' information use.

---

*Program Planning and Budgeting.

Are there any ways in which these obstacles might be overcome?

There are devices which, in theory at least, would correct one problem or the other. Perhaps the most obvious approach would be to change the constraints on "production" in education, so that schools are rewarded in proportion to the value they add to students' performance. Several variants of this notion have recently become popular, including merit pay for teachers and performance contracting for schools. Such schemes create new supply standards based on predetermined performance criteria. In theory, at least, these bureaucratically-established criteria become the "demand" which producers would seek to satisfy.

Would such arrangements increase the demand for information by educational producers? Since the notion of performance rewards implies one (or perhaps a very few) measurable criteria of performance, all suppliers would be interested in the same sort of information on the educational "production" process. Some of this might arise from a system of social indicators. Much of the information demand in a system of performance rewards, however, would probably involve technical innovation. Here social accounting systems would not be much help.

The really important question, however, is whether a system of performance rewards would strengthen the position of consumers and clients (that is, parents) with respect to information about schooling. One hypothesis is that such a regulated market system would work, and producers would disseminate information freely in their efforts to compete for clients. But prior experience suggests that it is probably more reasonable to suppose that producers would collaborate to minimize competition by maintaining performance parity and fixed shares of the market. They might also provide consumers with deficient or misleading information. After all, if a direct link between performance and reward were established, the most sensible course for producers would be to set some acceptable level of performance that most could meet, and close off further competition. Indeed, even if the producers did not take this tack, many students and parents might. Greater productive "efficiency," after all, would almost surely come out of the students' skins.

The likelihood, then, is that consumers and clients would be faced with many of the same problems they confront in "free" markets elsewhere. Much of the competition there lies not among firms to provide goods and services more efficiently, but between consumers and firms to find out what, if any, real differences exist among products, and what fair value is. This would not particularly help consumers of schooling. It certainly would not provide a situation in which information systems would give them appreciably more leverage in bargaining with educational producers.

In theory, of course, these combinations of producers against free markets would not occur. Ideally, in a performance reward system, social indicators would counterbalance the producers' tendency toward stasis. Information on the relative standing of schools' inputs and performance, for example—which could be easily incorporated within a social accounting scheme—would allow effective action against underperforming schools.

The difficulty, however, lies precisely here. Who would take action? The heart of the performance reward idea is that "market forces"—that is, the pre-established demand criterion—would compel producers to redress their own poor performance. Consumers would therefore really be quite passive. The important transactions would take place between producers and whatever agency collected information on their performance and disbursed the performance rewards. Therefore, even if we hold apart problems of fraud, price-fixing, and deceptive advertising, a performance reward system would not directly involve the clients and consumers of education. Indeed, the greater technical complexity might further separate them from the decisions.

At bottom, then, performance reward systems are really a form of government regulation, in which fiscal constraints replace bureaucratic or political punishment as the enforcement mechanism. Are there other schemes which might avoid some of the pitfalls of performance rewards?

One possibility is establishing countervailing centers of bureaucratic power, which might improve the schools' use of information and serve as consumer protection mechanisms. One way to do this would be to create sizeable independent staffs for local boards of education. They would have a mandate to monitor the effectiveness and efficiency of the existing enterprise, and an obligation to publish regular reports rating schools and services. Another would be to establish regional or statewide units with the same mission. Another would be to offer public subsidies for independent citizens research agencies, akin to the private government research bureaus that have been common in the larger cities since the salad days of the Progressive movement.

Such schemes would have several plain advantages over performance rewards. First, performance rewards involve a unitary output standard (or at best two or three standards), but astonishingly little is known about the "important" outcomes of schooling. Achievement test scores seem to have no direct impact on performance later in life. There may be an indirect effect, but we are not sure what it is. What is worse, even if we knew what was "important," people (and population subgroups) would differ in the degree to which they regarded the important outcomes as valuable. Any system of performance rewards, then, would be arbitrary at best, and perhaps mistaken. Competing bureaucracies, however, could deal with a variety of outcomes, at different times, and with different emphases. Their purpose would not be to insure performance in some mechanically rigorous sense, but to create incentives and constraints by political and administrative pressure.

Such agencies would almost surely utilize the products of a system of social accounting. In fact, they might become one of the chief consumers and interpreters of the new information. If the information were national or regional in scope—as it almost certainly would be—such agencies might gather similar data at the state or local level. More important, the information might actually be of assistance to consumers. One could argue, at least, that such agencies would avoid the consumer exclusion inherent in the performance reward schemes because they would rapidly discover that mere publicity was not enough. Headlines on Monday rarely produce change on Tuesday. The agencies might therefore try to generate support among parents by assisting established consumer groups or encouraging the creation of new ones. This would surely increase the availability of information to consumers, and it might even have some impact on schools. In theory, then, consumer groups would have a symbiotic political relationship with these new regulatory agencies: the former would have power, but not much capacity to gather or process information, and the latter would have the information capacity, but not the power to turn its product to political advantage.

The trouble with the theory is not that it is incorrect—but that consumers are by no means the only available constituency. Even the rosiest review of independent regulatory agencies reveals that they tend to be staffed by people from the professions or enterprises they are supposed to oversee, who act as though these professions and enterprises were their most important constituency. This has certainly been the case with the state school agencies and school accreditation groups, and it even seems to be true of the more independent federal agencies. Apparently the only way this tendency can be minimized is to mobilize consumer groups and force the regulatory agencies to work more effectively.

Thus we seem to have come full circle. Establishing countervailing bureaucratic power might impel the schools to make better use of information and to improve their performance, and it might help consumers use information about education. But these things seem unlikely to happen unless the new agencies organize the consumers. All past experience with such agencies indicates that they would be more likely to pay attention to the schools than to the schools' clients. This tendency would only be reversed if consumers forced the agencies to behave otherwise by applying political pressure.

The missing ingredient, then, is consumer power. Performance rewards and countervailing bureaucracies would exclude clients by confining the regulation process to competing centers of bureaucratic power. Or, to put it another way, both schemes would substitute government standards for consumer preferences. While parents and children would remain the clients of educational institutions, they could not have much influence on producers by changing preferences or switching brands. Such power would be vested in government hands.

In effect, although both schemes seek to make schools more responsive and to create better information use, both might founder on their exclusion of consumer interests. This should be no surprise. These schemes propose to affect schools' behavior by constraining the terms under which schooling is supplied, but students and parents are not suppliers of education. Thus, government regulation of supply leaves them as passive bystanders to the process of their protection. It offers them neither incentives nor new avenues for informing themselves, or for policing the action of various government agencies.

No amount of government regulation could remedy this difficulty. The only way to increase either the consumers' ability to utilize information, or their power to compel public agencies to do so, is to increase consumer power. This would involve altering the constraints on demand, rather than seeking to further regulate supply. To be precise, it would require that parents could choose among schools.

There are a variety of mechanisms that would allow consumer choice. One would be permitting small groups to receive state subsidies if they wished to establish public schools; another would be elimination of zoning requirements for public schools and allowing parents to choose among them freely; a third would be to permit community or other groups to subcontract with the existing school systems to operate all or part of a school. Finally, parents could be given tuition vouchers, which would allow them either to choose among existing schools or to join with other parents to form new enterprises. Vouchers are probably the most effective device.

Any of these would be rather a large step. While government regulation of the supply of public goods is no novelty, consumer choice among public service producers is almost unheard of. In my view, however, it would be most likely to sharply change the schools' pattern of information use. For one thing, tuition vouchers would provide a simple and direct incentive for schools to do the things they promise because the vouchers would give consumers the power to go elsewhere. This is the same sort of incentive as performance rewards (that is money), but the consumers, not the state, would control the incentives. As a result, they would be much better situated, and more motivated, to demand information on the schools' performance.

This is not to say, of course, that there would be no tension between consumers and producers, or that producers would not try to control information or present it in the most advantageous terms. It means only that consumers would have a weapon that would give them some bargaining power with schools, and some reason to combine to secure good information. That is, it would tend to encourage the formation of consumer protection groups, since parents exercising choice among products would desire some independent assessment of the alternatives. A review of consumer behavior in other markets, however, suggests that this would be far from a universal phenomenon.

Finally, it is worth noticing that vouchers would work even where performance was measured in different ways by different schools. Unlike the performance reward schemes, parent choice would require only that schools do the things they promised. Although some of these things might elude a purely quantitative system of social indicators, many would not. The information system required would be more complex than in a performance reward scheme, but that would hardly discourage the advocates of social accounting. More important, the information collected would be of interest to both consumers and producers.

Tuition vouchers would not produce perfect information use or anything approaching that. Individual consumers are always at a disadvantage when they confront large, organized enterprises, and this would be no exception. Indeed, there is good reason to believe that even with a system of client choice, it would be essential also to have an independent government agency to collect, process, and publish information on schools: what they promised and how delivered. What is more, all the schemes I have discussed leave untouched the problems of differences among population subgroups in the capacity to use information. While parent choice would help most in this respect—because it would encourage consumer unions, rather than leave individuals isolated—differences would surely persist.

Given these problems, however, empowering consumers seems to hold the greatest promise. It would be most likely to increase citizens' power to utilize information, and their ability to compel schools to do the same. It would, in a word, improve both the schools' demand for information, and the consumers' ability to utilize it.

Conclusions

This paper has been a preliminary foray into a complicated area—the political barriers to social intelligence in education. My argument is that the main obstacle to social accounting is that schools are not organized to utilize such information, and that at present consumers have no way to change this. Of several possible remedies, the most promising seems to be consumer choice among schools. This would provide a substantial incentive for both schools and consumers to seek and utilize information.

Of course, this is very abstract. Loosening the constraints on consumer choice might also affect racial segregation, economic discrimination, and church-state relations. Avoiding problems in these areas might require some constraints on consumer choice, and one does not know what effect this would have on information use. It also is possible that bureaucratic regulation schemes would work much more effectively than I have suggested, or that there are better ways to create countervailing power than those I mentioned.

In each case, it would be worth the effort to find out. And perhaps the most important point we can emphasize is the need to experiment with new institutional models. Were different approaches to constraining supply and unconstraining demand tested, we might learn a good deal about the behavior of schools and their clients under changed conditions.

The reasons for such experiments are far from trivial. We live in a society that has always officially subscribed to the notion that reason is regularly and successfully applied to public affairs. Indeed, the last decade has seen a rising interest in the application of systematic intelligence to society. Studies of the future, of social indicators, of PPBS, of evaluation are only a few manifestations of this. I have no doubt that the next decade or two will see an enormous increase in social information, but our capacity to manage and apply this information lags dangerously. This is a problem of social and political—not machine—technology. Our invention of ways to produce and process information accelerates, but our ability to digest and utilize it does not. Many resources are committed to the technology of gathering and processing information, but few to its social utilization.

Thus, while I am an avid advocate of more and better social intelligence, experimentation with new organizational forms seems much more important. I say this because the inventors and interpreters of information systems have a responsibility beyond simply creating them. Information, after all, is social, and the rationale for its existence is its social utility. If there is good reason to believe that new information will not be very useful because it will not be used, it would be perverse to do no more than continue to generate it. The more sensible course would be to devise ways to increase the chances for its utilization.

This course would not be easy. It would require efforts to understand and overcome the schools' resistance to the application of organized intelligence and their resistance to their clients' preferences. But not to do so may in the long run be worse. After all, what better way could be devised to undermine the case for social intelligence than to create it in situations where there is little hope it will have any use?

REFERENCES

1. Coleman, James S., and others. *Equality of educational opportunity.* Washington, D.C.: Department of Health, Education and Welfare, Office of Education, 1966.

2. Stephens, James M. *The process of schooling: a psychological examination.* New York: Holt, Rinehart and Winston, 1967, Chapter 7.

3. Flanagan, John C., and others. *Project* TALENT. Pittsburgh, Penna.: University of Pittsburgh, 1964.

4. Wholey, John S., and others. *Federal evaluation policy: social program evaluation by federal agencies.* New York: Taplinger, 1970.

Ethical and Legal Aspects of the Collection and Use of Educational Information

DAVID A. GOSLIN
Russell Sage Foundation

In March 1970, Russell Sage Foundation released the report of a conference it had sponsored on the ethical and legal aspects of school record keeping. Entitled *Guidelines for the collection, maintenance and dissemination of pupil records* (1), the report received considerable attention in the press at the time of its release and subsequently has created great interest among parent groups, school administrators, researchers, and others concerned with our nation's schools. Copies of the report have been widely distributed by the Foundation with the cooperation of the American Association of School Administrators, the National Association of Secondary School Principals, the National School Boards Association, the American Personnel and Guidance Association, and many local organizations having an interest in this problem. To date, nearly 100,000 copies of the report have been distributed, and requests for copies are still coming in at the rate of several hundred each week.

The report made headlines by calling attention to the absence, in most school systems, of any clearly defined and systematically implemented policies regarding uses of information about pupils, the conditions under which such information is collected, and who may have access to it. A number of examples of potential (not actual) abuse were cited, and in the preamble to the recommendations the conferees stated that "It is our opinion that these deficiencies in record-keeping policies, taken together, constitute a serious threat to individual privacy in the United States." The intended meaning of this statement is that present practices create conditions which make possible intrusions on the privacy of pupils and their parents—not that such intrusions occur in all or, indeed, even in very many cases.

Despite the headlines, however, most of the report was devoted to the presentation of explicit guidelines for the development of record-keeping policies in schools. Not intended as a muckraking document, the report was designed to be helpful to school personnel, parent groups, and others by providing them with concrete bases for discussion of the issues. Among the major recommendations of the report were the following:

> No data, including standardized tests, should be collected about pupils without the informed consent of parents, and in some cases, the child. (Specific procedures for obtaining such consent were proposed, with full attention to the administrative burdens already being borne by schools. For example, a distinction between individual consent and representational consent was proposed—and conditions specified where each would be adequate. The report even includes sample permission forms and a series of hypothetical cases to help school officials and others interpret the recommendations.)
>
> Schools should establish procedures to verify the accuracy of data contained in pupil records and for periodically destroying information no longer needed.
>
> Parents should have full access to, and the right to challenge the accuracy of, data on their children, and no persons other than specified school officials and parents should have access to pupil data without either subpoena power or parental and pupil permission.

There is a great deal more to the report than these general statements can convey, but rather than spend time reading detailed recommendations, I should like to concentrate on the issues which led the Foundation to convene its conference, and on some of the reactions to the report.

Background

The conference that produced this report had its origins in several different activities and concerns which have been a major focus of Foundation interest during the last eight or nine years. As most educators know, throughout this period Russell Sage Foundation has been supporting a program of research on the social consequences of standardized testing in American society. Several research reports have resulted from this endeavor, including a survey of teacher attitudes toward and uses of tests by Goslin (2), a recent volume on American attitudes towards intelligence by Brim and others (3), and a survey of record-keeping practices in schools by Goslin and Bordier (4). Forthcoming reports include a study of the test publishing industry and a survey of testing in business and industry. Among the many issues identified by these studies were the right of a pupil or his parents to have access to test scores compiled by schools, the possible impact of such information on the pupil, and the school's responsibilities with regard to information about pupils contained in its records.

Perhaps most significant of all our findings was confirmation of the fact that enormous variability exists in the use that is made of tests by schools and by individuals within schools. No one, including teachers or counselors themselves, appears to know, for example, how much reliance is placed on test scores in making decisions about pupils, evaluating their capabilities, and adapting teaching techniques to fit the needs of individual pupils. It is very clear, however, that schools currently collect and maintain a great deal of information about pupils (and their parents) in their record files. It is equally clear from our research that the accuracy of this information, what use is made of it, and who is permitted access to it are determined almost by chance in many systems.

Another factor which led the Foundation to convene its conference on record-keeping practices in schools was the growing concern in American society with the protection of individual privacy. The increasing size and consequent bureaucratization of all major institutions in the society, including the school, coupled with advances in computer technology and the electronics field have raised important questions about what must be done to preserve the right of individuals to personal privacy while at the same time recognizing the legitimate claims of society. As Oscar Ruebhausen stated it in the preface to the report, "Modern science has introduced a new dimension into the issues of privacy. There was a time when among the strongest allies of privacy were the inefficiency of man, the fallibility of his memory, and the healing compassion that accompanied both the passing of time and the warmth of human recollection. These allies are now being put to rout. Technology has given us the capacity to record faithfully, to maintain permanently, to retrieve promptly, and to communicate both widely and instantly, in authentic sound or pictures or in simple written records, any act or event or data of our choice" (1).

Record keeping, of course, in one form or another is an integral part of the educational process. At the simplest level, an educational record describes changes taking place in individuals that may be attributed, at least in part, to their participation in the teaching-learning experience. Since change (learning) is presumed to be the primary goal of education, the record of such change provides a measure of the effectiveness of the educational process as a whole as well as of the performance of its participants, principally teachers and students.

From the beginnings of human society, teachers have no doubt kept track of the performance of their pupils. Effective teaching, no matter how informal, requires that the teacher have some idea of what his pupil knows and does not know, how quickly he is able to grasp new ideas or acquire new skills, and what kinds of learning are especially easy or difficult for him. Similarly, the student's motivation to continue to engage in the educational process is no doubt related to his perception that he is making progress, a perception facilitated by the maintenance of records. Moreover, the record of an individual's performance in learning situations long has been used as an important indicator of his capacity either to handle tasks that require the utilization of previously acquired skills or knowledge or to engage in new learning.

Educational records may be expected to reflect accurately characteristics of the educational process. Simple educational systems, typified at the extreme by a one-to-one teacher-pupil relationship focused on the transmission of a single set of interrelated concepts or skills (for example, a father and his apprentice son), are characterized by highly personalized and informal record-keeping techniques: a diary, collections of work done at various stages in the process, or even individual recollections corroborated by the observations of others. Complex systems, on the other hand, necessitate more complex record-keeping procedures.

The United States currently possesses the most highly developed and complex educational system any society has ever created. A great many changes have taken place in the characteristics of educational institutions in America during the past 50 years. These changes have been the result of (and in turn have contributed to) broad changes occurring throughout the society: technological advances; demographic changes, including urbanization and suburbanization; shifts in political and religious attitudes and values; and so on. As a whole these major alterations in the society have perhaps had less of an impact than some observers have claimed on the basic *conduct* of education (that is, what is taught and how it is taught); however, they have resulted in radical alterations in the *structure* of our educational institutions. The most important of these structural changes are related to increases in the size and complexity of the educational enterprise, both with respect to units within the system and the system as a whole. Put simply, a much larger proportion of a larger population is attending bigger schools that are part of bigger school systems and for a longer period of time. Concomitantly, the society's investment in education has increased substantially at all levels; schools and colleges have become more specialized; the range of options open to individuals with respect to the educational experiences available to them has expanded rapidly; and, finally, the conduct of education has become a major focus of concern to many segments of the population that formerly took for granted what went on or did not go on in schools.

Even more significant, as the society's interest and investment in its educational systems have grown, schools have increasingly been charged by their constituency with responsibility for making sure that students work up to their capacity, for overcoming deficits created by cultural deprivation during the preschool years, and for helping pupils choose careers appropriate to their skills and interests. No longer do we conceive of the school simply as an institution offering certain kinds of training and knowledge to those with the interest and energy to learn. The school is expected to take positive action to motivate pupils, to understand their problems, and to remedy their deficiencies, both academic and personal. The school is put in the position of seeking and trying to make use of more information about its pupils. In addition to keeping a record of how much Johnny has learned, the school must also try to find out why Johnny *didn't* learn, how much Johnny *should* learn, and what the school can do to help Johnny learn more, if it is to do what is expected of it.

Ethical, Legal, and Social Issues

What kinds of ethical, legal, and social issues are generated by these developments? In answering this question it is necessary to take into account differences in the kinds of information maintained by schools.

School records typically contain two kinds of information about pupils. The first is the record of their activities and performance in school. It is comprised of the attendance record, systematic teacher observations and evaluations (specifically, grades), reports from counselors and other school personnel concerning their behavior outside the classroom, achievement test scores, a listing of extracurricular activities, and so on. The second type of data concerns the pupil's background, characteristics of his family, his out-of-school activities, and basic intellectual and personal qualities, including health, intellectual capacities, and personality dispositions. This distinction is an important one, since many of the ethical, legal, and social issues raised by current record-keeping practices have greater relevance to one or the other of these categories of information. Few persons, for example, would question whether it is legitimate and appropriate for the school to maintain records containing information of the first type. Clearly schools must have a record of the past performance of children in order to do their job.

Collection and maintenance of the second type of information, on the other hand, poses the issue of the grounds on which the school may legitimately ask pupils (or their families) to reveal facts about themselves that may not directly be related to performance in school. Very important values in American society suggest that it is a basic right of individuals to decide to whom and under what conditions they will make available to others information about themselves. Correlative to this point, however, is the fact that participation in the society carries with it certain obligations and responsibilities. Further, the right of groups to demand information from those who aspire to enjoy the privileges of group membership is clearly understood. Thus, no one is likely to object to being given a driving test before being permitted to operate a motor vehicle. Similarly, few people object to the requirement that they must take an entrance test in order to gain admission to a university or college. In each of these cases, the right of a group, in this case the school, to information that is necessary to achieve its stated objectives and goals has been established beyond question. However, some important considerations remain.

First, on what basis do we decide that certain kinds of information are necessary in order for the school to perform its function? As we have pointed out, school officials take the position that in order to do what is expected of them by the society, they must have a great deal of information about pupils. Measurement of intellectual capacity (for example, IQ testing) is defended on the grounds that the school's resources are limited and that pupils with different abilities have different educational needs. Measurement and recording of personality characteristics is justified by pointing out that understanding and compensating for deficiencies in performance, disruptive behavior, or other problems, requires knowledge of the "whole child," not just his intellectual capacities. Similarly, collection of data on family background makes it possible for the school to anticipate educational needs and deficiencies. Although it is doubtful that schools would cease to function if they did not have access to such information, a strong case can be made that more information about pupils not only makes the school's task easier, but also can help the school do a more effective job.

Second, having once established the criteria for assessing necessity (which we do not claim to have done), under what conditions does a group have the right to ask aspiring members for information that is clearly *unnecessary* to the purposes and goals of the group? To answer this question, it is necessary to make a distinction between public and private groups. It seems reasonable to assert that a private group has the right to ask applicants for membership anything it wants to ask them, relevant or irrelevant. In this case, it is up to the applicant to decide whether he wishes to reveal this information. In the case of a group supported by society as a whole, including all of the potential applicants to the group, this is a more difficult question. Would it be legitimate, for example, for the state to ask individuals to reveal information about their sex lives as a requirement for obtaining a driver's license? Most of us would object to such a requirement on the grounds that it represents an invasion of our privacy that is not justified by the service being rendered. Just such objections are being raised to the use of personality and IQ tests in schools, as well as the maintenance of a variety of other information ranging from anecdotal to clinical observations and family background data. These objections require us to consider much more carefully the need of schools for such data, their validity, the uses to which they will be put, and the conditions under which the school may legitimately collect them.

To sharpen our thinking on these points, let us suppose that children (or their parents) exercised the right to refuse to take any tests given by the school. If a child refused to participate in classroom tests it would, in turn, be legitimate for the school to refuse to promote him to the next higher grade. Few would argue that schools should not have the right to require pupils to demonstrate their *proficiency in* *school subjects* before according them advanced status. If this happened, however, it would be the child's (or his parents') decision. On the other hand, what if the child refused to take an IQ or personality test given by the school, or to fill out the information form that ascertains his family background? Could the school legitimately fail to admit him or promote him in this instance, assuming he was meeting school standards for proficiency in his daily work? Does the school need this information in order to evaluate his performance in school?

Access

Once information of either the first or second type (or both) has been collected and entered into school records, the question of access to this information must be faced. Both the rights of certain individuals (such as school personnel) to make use of this information and the rights of the pupil (and his parents) to be protected from indiscriminate use of the information by nonschool personnel are involved. In addition, the right of the pupil or his parents to know what information the school possesses about the pupil must be considered. In the latter case, at least one court has established the legal right of a parent to inspect his child's permanent record, despite the fact that our data show this practice to be contrary to the policies of most school systems. Even assuming that school systems were to accept this judgment at face value, however, the legal definition of the permanent record requires further clarification, especially if school systems were to attempt to avoid revealing certain kinds of information (for example, test scores, clinical evaluations, and the like) to parents by claiming that it was not part of the permanent record. The rights of the pupil in the matter also require clarification. Does the pupil also have the right to know what is in his record? Does he, under any conditions, have the right to prevent his parents from knowing what is in it?

Access of all nonschool personnel and some school personnel (such as teachers not responsible for a pupil, the research staff, and others) to pupil records is another very difficult issue. The major point of contention involves specification of the conditions under which data gathered for one purpose (namely, education) may be used for some other purpose without the consent of the individual (or his parents) from whom the information was collected. It is apparent from our questionnaire responses that schools frequently permit access to pupil records by a variety of outside agencies and individuals, in most cases, we suspect, without obtaining parental permission. Regardless of the strictness of school policy regarding access by outside agencies, all pupil records presently are subpoenable by the courts themselves. In most states counselors and school psychologists do not yet enjoy the protection from the law accorded lawyers and doctors with respect to privileged communications.

The Report and Responses to It

These were some of the issues that led to our report. Initial responses to it on the part of some school officials (many of whom were called by reporters before they could examine the report) were predictably defensive. While agreeing in general with the principles it advanced, they denied that any serious violations of individual privacy could result from current practices and went on to criticise the *Guidelines* for imposing unnecessary administrative burdens on schools. Equally defensive and even more upset have been many researchers who saw our report as threatening to raise insuperable barriers to the conduct of many of their studies. The problem of data collected under conditions of anonymity from sufficiently large populations to make identification of individuals impossible is, we feel, of substantially lesser importance than the others raised by the *Guidelines*. Nevertheless, to deny that abuses can or do occur in schools, or on the part of researchers, is, unfortunately, to miss the intent of the report, which was to describe as clearly as possible what a reasonably adequate system for insuring the accuracy and confidentiality of pupil records might look like. Similarly, criticisms of the *Guidelines* on the grounds that their full implementation would create undue hardships for schools, or would prevent school personnel from doing their jobs, would appear to be self-defeating. As is clearly stated in the report, conference participants were fully aware of the difficulties that some of the recommendations might cause schools. The *Guidelines* were not presented, however, on an all-or-nothing basis, nor were they intended to be the last word on these issues. Instead, the report was designed to serve as the basis for an informed dialogue among parents, students, school officials, and other interested parties concerning the most reasonable means of correcting current deficiencies in record-keeping practices *without* unduly hampering conscientious administrators, guidance personnel, or even researchers.

Most school systems today are being confronted with increasing demands by their constituencies, both parents and students, to accord them a larger share of the responsibility for decisions affecting the way schools are run. Pressures to reverse the long trend toward greater specialization and professional responsibility for educating children have often resulted in defensiveness on the part of educators, both with respect to their competence and with regard to the complex organizational structures within which they have operated. To those who advocate radical reform of our educational institutions, the seemingly impenetrable bureaucracy of our school systems becomes a symbol of many of their faults.

In this context, the issues raised by the Russell Sage Foundation report take on an importance far greater than the question of how frequently current record-keeping practices actually jeopardize the privacy of students or their families. More significant is the question of whether school systems will be willing to draw back the curtain of secrecy which currently surrounds many of their activities and permit students and parents to participate as partners in the educational enterprise. To do so, of course, is to run the risk of increasing the intensity—at least in the short run—of criticisms of schools and of school personnel. But to persist in insisting that parents and pupils should leave all decisions in the hands of professional educators would, in the long run, appear to be far riskier—if not sheer folly.

Forthright and open discussions among all of the interested parties regarding the problem of school records and their management would appear to offer a major opportunity for schools to begin to restore the confidence of their constituents in their goodwill and integrity. Whatever specific policies might result from these deliberations, such a process should produce dividends in increased confidence and cooperation among parents, students, teachers, and administrators that would far outweigh possible added expense and administrative effort.

As important, then, as the *Guidelines* themselves is the process by which schools move toward their implementation. As stated in the report itself, "In keeping with increasing demands for participation by students, parents, and community leaders in the governance and rule making in the school, we urge that the very drawing up of such a code for the definition, operation, maintenance, and disposition of sensitive school records should be subject to student participation within the school and to various kinds of consultative referenda or clearance with key parent-teacher associations, community action groups, and professional associations within the community. The issuance, by administrative fiat, of a set of rules by the school system, carries with it the danger of insuring misunderstanding by the various populations whose trust and goodwill must be linked with the system if it is to operate with maximum effect."

It was with these various goals in mind that Russell Sage Foundation convened its conference on the ethical and legal aspects of record keeping in schools. It is the Foundation's hope that these recommendations will lead not only to improved procedures for the management of pupil records, but also to closer cooperation among pupils, parents, and the schools.

REFERENCES

1. *Guidelines for the collection, maintenance and dissemination of pupil records.* New York: Russell Sage Foundation, 1970.

2. Goslin, David A. *Teachers and testing.* New York: Russell Sage Foundation, 1967.

3. Brim, Orville G. Jr., Glass, David C., Neulinger, John, and Firestone, Ira. *American beliefs and attitudes about intelligence.* New York: Russell Sage Foundation, 1969.

4. Goslin, David A. and Bordier, Nancy. Record keeping in elementary and secondary schools. In S. Wheeler (Ed.), *On record: Files and dossiers in American life.* New York: Russell Sage Foundation, 1970.

Test Information as a Reinforcer of Negative Attitudes Toward Black Americans

ELIAS BLAKE JR.
Institute for Services to Education

As I have approached the particular point when I would get my opportunity to talk—sitting here, listening to the presentations preceding mine—I have had many different feelings. And I think I have felt today in many ways like a lot of people who are, maybe, 15 or 20 years younger than I am, particularly as the day progressed. That is, one wonders sometimes—to put it as they would put it—just what the hell is going on.

Maybe it's because, being black, I must see myself as being somehow at the center of a great many of these issues that have been discussed with great intellectual verve, with bon mots, with ripostes back and forth, and so on. But I am left with a funny feeling that this dialogue really isn't dealing with the *urgency* of the issues. We talk of all these alternatives among which, once we get all of the information in hand, once the computers finish grinding out their data, we will decide which we are going to follow, and *then*—only then—we are going to save the world.

I keep feeling that in the meantime—in the *meantime*, this very present meantime—all these things that disturb me and many of you are still going on. Though I would like to put a positive face on things in many instances, I must confess I feel less positive at this point in the day than I felt earlier.

I think essentially I'm most concerned about the fact that social scientists as a fraternity do not think enough about that word "social" in the science that they are dealing with and are supposed to represent. They must somehow remember that in the background of all these findings, conclusions, and generalizations lie *a* people, and *all* people, and somehow these professionals and their groups have got to start thinking about the kinds of things in which they will *not* become involved, as well as the kinds of things in which they will be involved.

I think the political climate today demands that we think a great deal about this sort of thing. Now let me address myself to the specifics of my topic.

Several issues will be discussed briefly and, I hope, provocatively about information systems that are based primarily on test data and its impact on groups whose performance reveals iniquities in their treatment by organized educational systems—kindergarten through college. A more general proposition I would like to explore is that as long as there are interlocking relationships between money—and money is involved in the testing movement—and status, cultural values and the use of tests, major alterations in this system are unlikely. This conference, in fact, may be evidence of the thesis. The wealth of the testing industry supports us handsomely to discuss problems created by the existence of the industry itself. When this happens on other fronts and in other fields, we are properly cynical and critical of such an intercorrelation; but somehow our own dependence within this, our system, is not viewed as a fatal flaw. We view ourselves, I suppose, as better than the regulatory agencies Ralph Nader castigates for being too sympathetic to the interests they regulate. Yet we must remember that we ourselves—this amorphous profession of researchers, teachers, counselors, and administrators—are about all there is for regulation in education.

Everyone will agree that we must have basic data on academic performance, cognitive skills, achievement or intellectual skills—or other appropriately neutral sounding assessment labels. Everyone will also agree that sensitive data on backgrounds and personal motivational traits are more dangerous, and that subjective entries are the most dangerous of all. But how, then, is the least sensitive data—that is, with the neutral labels—to be made more useful? How are the abuses of its use to be ended?

Alas, if the present scenario evolves, the industry will emerge clothed with the recommendations of more eminent scholars and authorities which can be used to safeguard the status quo, while the scoring machines and printers and computers spin merrily along. Though I would like to dwell on the larger issues—for example, whether the presence of all the test data has, in fact, been an advance over its absence—I will move on to the topic at hand.

The first point I'd like to make is that the achievement testing system has been a negative force in the attack on the educational problems of black Americans. There can be no doubt about this. At the same time, by revealing the existence of a gap in performance on these tests between black and white as *groups*, it has been very useful to advocates of equality of opportunity. However, I would argue that concentration on socioeconomic and demographic data has supported a raft of alibis and excuses and diverted attention from the main question. Our gaze has been diverted from the central question of the quality of interactions between teachers and students in individual classrooms. The problem, after all, is what happens between individuals in individual classrooms.

Meanwhile, the mass of this data showing lower scores for blacks shapes attitudes and expectations in a deep and pervasive manner. The essential uselessness of some major test-reporting formats for building instructional programs has been demonstrated—for example, age-grade norms and percentiles. The lack of substantive teaching data in these scores doesn't seem to have generated much reform in the testing industry. These essentially useless reports also act to depress efforts at reform, since it is extremely difficult to show significant or dramatic gains that will stand the scrutiny of sophisticated methodological critiques when they are used to measure change. Thus, countless social programs such as Headstart are undermined and become questionable due to the use of these indices. Thus, the future of the education of black Americans moves from strategy to strategy to strategy in search of the significant and replicable results. The net effect is that the absolutely vital commitment to long-term, longitudinal, sustained and persistent efforts leading from preschool education at least through high school graduation is being forever delayed, mainly because "research" based on this kind of test data is too ambiguous. The test data, then, are making a difficult political problem even more so.

It would be interesting to examine the career patterns of the researchers who argue over whether Coleman really found out anything. Though brave noises are made, it all seems to settle eventually on the usual tests as the criterion of significance. Then political figures use the statements of whoever surfaces out of the scholarly arguments with documentation for their particular brand of political solution. Certainly the lack of commitment to putting up the necessary money for long-term support at the local, state, and federal level is not caused by these arguments. However, by this means the lack of commitment is given the devastating support of this veneer of scholarship and research backing.

For example, I sat through a particularly depressing briefing on the analysis of test data on Title I of the Elementary and Secondary Education Act given by a staff member of ETS to the Office of Education staff people. Though the researcher bravely implied that the quality of what was happening between teacher and students was the real problem, the real issue, the data he presented, of course, did not deal substantively with what he had admitted was the most important issue. The technical arguments that followed that presentation were hot and heavy.

My concern, however, was a larger one: that nothing dealing with long-term effects—say, up to entry into high school or college—was being contemplated. At the very least, such thinking would begin to develop the climate of thinking for the long-term commitments of funds, and it would require some different criteria and research strategies.

The second point I would like to make is that the quantitative testing movement and its attendant industry have pushed social scientists ever more deeply into the powerful but doubtful world of mathematical models. Mathematics gains power by dropping out things to gain useful abstractions, and I am concerned by much that drops out in this process as used by the testing industry. There is something basically weak about the idea of basing one's fate on one-time, one-day marks on pieces of paper. I am sometimes astonished by our reliance on such a narrow series of responses. If you push any researcher on this issue, he will acknowledge the weakness. But always this issue fades into the background, and throughout education people act as if they really had a wide range of data on all possible relevant behaviors on which to base the kinds of decisions they make and the discussions we have.

Another questionable assumption that too many seem to have is that social science will one day approximate the physical and natural sciences in producing solutions to complex problems. This seemingly fine belief is becoming very harmful because of what it is doing to black Americans. For, based on this assumption, social scientists keep holding out promises that large-scale solutions will be found to the problems involving blacks, and this can be very destructive. They promise what they cannot deliver, even with the most comprehensive data banks. So much eminence, and prestige, and credentials, and intellectual power so long at work with so little success is dangerous. If they cannot succeed, one fears that the general conclusion will be that the subjects of all this attention are incurable.

It must be remembered also that only certain segments of American education are dysfunctional. The middle and upper middle class parts of the system work quite well. They do exactly what people want them to do—and you can find this out by going into any middle or upper class community in this country. You start tinkering with their system and you are going to be in trouble. They want things to stay precisely as they are. Their children go to college in very large numbers and they get out of college in very large numbers. They take very good jobs in very large numbers, and eventually they wind up in positions of policy making and decision in industry, business, education, and government. That is not a dysfunctional system for their purposes.

In contrast, in low-income neighborhoods a great deal of the mythology about why schools don't work is foisted upon people who do not have very much sophistication and need help with deteriorating schools. One interesting aspect of this has to do with this business of parental participation in the schools, which is supposed to influence the way children are educated, with a certain level of participation being a good thing. To me this is a myth. I don't know of any school system where we have much parental participation. What we have is those ladies for whom the PTA or home-school association is their pet project. They run it. They control it. They see that everything goes according to Hoyle.

We do not have very much parental participation in any schools unless some kind of issue is at stake, or at the beginning of the year, when they want to go into the school to observe the situation and see that it is satisfactory. Then they fade away, never to appear again unless there is some controversy.

I submit that PTA meetings represent this kind of lack of participation. It is true, and it's related to the corollary: The system works. Why fumble around with it? But when educators and social planners look at low-income areas and see some very serious problems in getting parental participation in the schools, what is their conclusion? They report that something is wrong with the parents. These parents don't understand, and so forth, and so on.

There are reasons why I would like parents to be involved in education, and they relate very much to the things David Cohen was talking about. I think they need to become more the watchdogs of the system, because, in fact, in the upper middle income areas many *are* watchdogs of the system. They evaluate the teachers through the comments of their children. If there are negative comments, they check up with the principal, to see that he knows what is going on, and then bring pressures on him about a questionable teacher. Any teacher who does not pass muster is phased out of the system, one way or another. They disappear, usually by transfer to less vigilant schools.

The social scientists, then, have a problem worthy of their most sophisticated skills in treating the dysfunctional segment of American education—that segment inhabited by blacks, other non-whites, and the poor. And I would suggest that maybe as social scientists we should reconsider and assume that what is found to be true in one school may have no applicability elsewhere, and that one must use what has been learned to start over again with a little more knowledge in a different school.

Since schools, as other social settings, are dynamic, fluid entities, the hope of large-scale generalizations may be a futile one. At the very least, social scientists should question severely whether they can ever duplicate the feats of sciences where controls can produce a fixed series of interrelations. For research on humans, we may never have that kind of control. The kinds of sophisticated manipulations of test data seem to imply that one day we will be able to know a great deal more about cause and effect relationships, for example.

In the foreseeable future, it is not at all clear that any level of the educational system will become responsive to or effective with black Americans. A lot of the comment I hear about higher education for everybody seems to me to be a separate issue from equality of educational opportunity. It may be that equality of educational opportunity from my point of view—that is, getting more blacks and other nonwhite minorities into all segments of higher education—may be related to this issue of more people generally wanting to go into higher education. But I think the two should be sharply differentiated, because what we are dealing with among black Americans and other nonwhites is an underrepresentation at all levels within the group, that is, high ability levels, medium ability levels, and low ability levels. That is a different kind of crisis situation and one on which we must move much faster than on this other issue of larger proportions of high school graduates going on to college.

Effective higher education for blacks in my view means in real world terms—in terms of staying in school, employability upon leaving school, the ability to enter into higher education whether Ivy League or community college. Such factors as these must be used as indices of the effectiveness of the system, rather than the height of entrance requirement scores or other arguable statistical indices.

I do not argue that information from tests is a cause of racist attitudes, but rather that it is a pervasive and convenient reinforcer of attitudes that are already negative toward black people. And the most ethically and rigorously handled data will continue to reinforce these negative attitudes. The most ethical action, then, might be to refuse to allow the use of tests where they contribute to such problems—as those in the case of intra-school groupings that are being used to cover up and carry out the deeper purpose of segregation and humiliation and subordination of black children.

It may be that tests should be handled like dangerous drugs, requiring both specification of their use and prosecution for their misuse. That is, maybe we should develop legislation governing the use of tests, so that people could be prosecuted for using tests in ways they should not be used.

I guess I would like the social scientists to put some teeth into usage requirements, or withhold the tests from all the people who are going to misuse them.

I was greatly disturbed this morning by one man who said, in effect, the test makers knew what they were doing and they did it right; it was all those other people who messed up. Then why didn't test makers put a moratorium on their sales and say, "We are not going to give you any more of these things until you learn to use them correctly"? But that gets into the problem of money, and there is a lot of money involved here. If there were more controls on the use of tests, maybe I would feel more comfortable with the people who run the testing industry and disclaim responsibility for their misuse.

My fourth point is this: The fact that so little is known about the relationships between cultural content and performance skills raises the issue of the value orientation of most tests. Today, the new questions raised about black cultural values bring this problem into dramatic focus. For there can be no doubt that much of the content of tests which are supposed to provide a demonstration of "culture-free" skills (such as reasoning and drawing inferences) has been alien to black Americans. Supposedly it makes no difference what content is used within a test if all the data and information are given for solving a problem. Supposedly it all reduces to simple reading ability and, say, the drawing of inferences from what one reads. There can be, the testing apologists repeatedly tell us, no questions of bias in such an approach.

More perceptive observers, however, feel that it goes deeper than this, that very complex and often very different cultural familiarities are involved. And these differences in seeing, hearing, feeling, and thinking on the part of black youth may be affecting their performance.

Let me give you one good example of what I'm talking about out of our own experience today—the joint experience we've shared of this conference. I was not particularly amused this morning by Dr. Fritz Machlup's use of chamber music and rock and roll as an analogy for "higher" and "lower" education. The implication is that fine chamber music is being threatened by the tyranny of rock and roll. In my view, it is more the other way around, with those in the academy who support and enjoy chamber music controlling the cultural apparatus so as to suppress or denigrate other kinds of good music in the society. What I was hearing while this audience was laughing robustly at Dr. Machlup's example was an insensitivity to the fact that rock and roll itself is a derivation of a more authentic music from the black community. The authentic music, rhythm and blues, and its performers are threatened both by the new commercialism of rock and roll based on white performers and the continuing snobbery of the academy.

Rhythm and blues and jazz have their great virtuosos, and they don't eat. New York City is full of them. Derived from and original to this society, this music is deeply rooted in our society and has dominated the popular culture for 50 years. But academicians, of course, don't find this to be an issue of any moment. Nor can most of them respect the culture out of which the music came. So I get very disturbed about higher education and lower education being described in this way.

The point I'm making generally, about values and testing, is that the relationships between the tests and the standard curriculums are likely to suppress rapid social change. Too many people worry about such things as the SAT and the Graduate Record Examinations and their value as entrance criteria or as things that set in motion more generally debilitating expectations. Where there is strong interest in a specialized curriculum, as with some black youth, another handicap is built into the assessment system. Their legitimate and scholarly pursuits in jazz or black history, for example, can wreak havoc with their performance on the standard tests. Then for the conservative observer their spotty performance becomes proof positive of the so-called "soul" or "cornbread" quality of their academic work, rather than a commentary on the different emphasis they choose to pursue. What is required is a wider variety of tests—and a wider variety of skills and interests within tests—but this need runs head-on into conflict with the standardizing needs of the testing industry for mass adoptions.

The last point I would like to raise is this: Discussions about testing seldom approach it in an economic, or profit or loss sense, but such an approach might prove highly suggestive in relation to questions of needful social change. For example, what would an economic analysis show about decision making on the policy level in testing? What kinds of decisions cannot be made without seriously damaging the anticipated income of the testing industry? What are the marketing techniques, and how much are they concerned with the proper uses of tests? What approaches to the uses of tests and test data might cause major retrenchments in the industry?

As one who watches interlocking economic forces create either opposition or indifference to his aspirations for social change, I would be interested in what these economic forces are in testing. Who dominates the markets, and why? What ancillary education professionals in schools and colleges are dependent on testing? Researchers? Counselors? Guidance personnel? What is their relationship to testing companies, and how much are they involved in decisions about the uses of tests and about changes of tests? What is the decision-making process for the adoption of a test in a state or in a major school system? How do competitors compete for these adoptions?

Out of questions of this kind might come some very useful new perspectives on information systems, their growth, and their control.

And last, may I make this observation out of a very deep concern. If the current climate in our society continues, it is likely that some of the implications of recent studies and speculations about racially determined genetic pools, and also the proposals on early identification of delinquents, will resurface and they may find support for field trials. I am concerned, then, that social scientists may forget that the humanity of at least one group of the people in the society is at stake, and that this is not simply the high-minded pursuit of purely scientific answers.

What continues to disturb me about the social science fraternity is that they continue to provide prestigious platforms for those who would reopen the question of racial inferiority as a subject that "scientific" data from the tests might clarify. I am sorely concerned with this, because I think social scientists are very naive about what they are doing in relation to their society, not as it exists in books, but in fact. They are not sufficiently alert to how their studies may be used for political purposes. They say: We must search after truth. We have to do what we do, and let the facts fall where they may—as if that was all there was to it.

I think there must be a greater dialogue on this particular issue. Because of the particular political climate we now have, and because of the way research dealing with social problems is being used in political circles, I think the social scientists through their societies—the fraternity itself—must do something about this particular issue.

The fact that the view of which I'm speaking has resurfaced in social science in our time, and the fact that its advocates can find all kinds of prestigious platforms frightens me; and it does not encourage me as I view the possible future effect of tests on black Americans.

Discussion

JAMES J. GALLAGHER
University of North Carolina

I think both Dr. Blake and Dr. Goslin were raising the essential ethical issues that Carl Rogers once put succinctly when he asked, "Should we do everything to people that we know how to do to people?" We still are struggling with the answer to that particular question, and that certainly was the focus of the "rights of the individual versus the rights of society" issue that Dr. Goslin posed.

Let me first discuss Dr. Cohen's suggestions, because he was raising a different kind of question: Are information systems, in fact, a change agent? He answered: "No, not by themselves." And I would subscribe to that answer.

The Program Planning and Budgeting System was mentioned as one type of a possible information system. I think it was George Bernard Shaw who, when asked what he thought of Christianity, said that he thought it was an interesting idea and hoped someone would try it out sometime. That's the attitude I would have to hold toward the Program Planning and Budgeting System. As it has been used in the Department of Health, Education and Welfare, where I served for three years, PPBS was seen as another information system rather than as a way of life. It *is* a way of life, and that is something that neither the federal nor state agencies are ready to accept at this time.

Dr. Cohen asked, "Is there really an informational gathering system at the state and local level in the educational system?" The answer, of course, is, "No, but there are a lot of people working very hard to try to establish one."

In his emphasis on the usefulness of consumer demand as a change agent, I think Cohen has made a partial diagnosis. The fundamental problem with voucher systems lies not in the problems of desegregation, or aid to parochial schools, or any of these other issues. The fundamental problem rests, I believe, in the incorrect assumption that the failure of schools to change or improve lies in the failure of will or inadequate motivation on the part of school personnel to change. The basic concept seems to be that, if there were a sufficiently large carrot or whip they *would* change, and the voucher system would provide that motivating source.

I would remind you that we do have a voucher system in another area of our society, in the delivery of health services. Each of us carries the monetary power to go to the physician of our choice, to choose between them, to figure out which one is better. But the delivery of health services in this country is still not one of our more striking accomplishments. The free enterprise nature of consumer demands has not encouraged an effective system in this instance, and there seems to be little reason why the delivery of educational services would be improved by the existing voucher proposals.

I have an alternative hypothesis that fits the data better. The problem of change in a complex organization is almost always a *systems problem*, rather than a people problem. We refer routinely to the American educational system, but there is no such thing. The American educational enterprise does not fit any definition of "system" that you ever saw, or that I ever saw. There is a collection of 20,000 relatively independent school districts out there, each governed by its own board and influenced very slightly by states, and very, very slightly by federal actions in education. The decision making remains basically at the local level, and the relationship between the service units at the local level and the support services which are really necessary for quality education lies clearly beyond the control of the local decision maker, or the school superintendent.

Four major dimensions of support systems are crucial in effecting educational change. They are manpower analysis and training, research and development, communication and planning, and evaluation. The local decision maker plays a very limited role in the manpower analysis and training and exerts very little influence on training institutions or agencies that provide training funds.

In the area of research and development, he is similarly limited. I can't build a DC-9 in my garage, and the local teacher can't build a new science curriculum integrating biology, chemistry, and physics. What they can do is respond effectively to programs that have been developed elsewhere. They can insert local variations, but they certainly can't produce the original program and, consequently, this key development is not under the control of the local administrator, either.

Discussion

There is no communication or transportation system to move new ideas, new concepts, new procedures in education from one place to another. If you have a great educational idea in Denver, Colorado, how do you get it to Miami, Florida? What's the standard system by which you move an educational practice from Winston-Salem, North Carolina, to Utica, New York? There is no standard procedure because there is no communication or transportation system of any merit. Even what there is now does not fall under the control of a local administrator.

Finally, there are few attempts at long-range plans or budgeting of resources to attack major issues in education, and these are not under the control of the local administrator. There has been only the beginnings of this kind of planning for systems at the federal and state levels. So the support systems crucial to the development of quality education are *not* under the control of the local administrator. Therefore, either giving him a carrot in terms of a voucher, or a whip in terms of withholding the voucher, is not really going to take care of the problem. Only by establishing these major support systems, plus systematic planning at a regional, state, and federal level, will there be a reasonable chance for continuous improvement in education.

In terms of the rights of individuals versus the rights of society, I think that most of the conflicts have been decided recently in the direction of society. As we get into an increasingly interdependent mode in our society, more decisions will go in this direction. Goslin's distinction between cognition, or academic, kinds of information versus personality kinds of information is not terribly useful. The goals of the schools have been broadened to include moral and attitudinal as well as academic objectives. It is in the nature of schools that they will need to collect attitude and personality data.

Testing is merely a special case within a general case. Any of you who know teachers who have given the familiar assignment "What did you do last summer?" to their students may recall the horrified look on a teacher's face as he read the essays, which often tell a great deal more about the family life and style of the youngsters than the teacher wants to know. What we need is much more clearly defined rules of confidentiality of information than we have had in the schools. The doctor and lawyer and psychologist keep personal information confidential. The educator must do likewise.

One basic freedom was taken away some time ago—the individual freedom of a parent to decide whether his youngster *should* go to school. We have had compulsory schooling for some time now, and we do that on the basis of a value decision: The child has a right to an education. I think we are wasting our time by asking, "Are we going to collect information, or aren't we?" We obviously *are* going to collect information. The problem is: How can we use our energies to protect the privacy of the child and parents in those dimensions as effectively as possible?

There are various organizational methods of consumer control and review that should be instituted with information collected by the schools. There should be a public accountability of the institution to its clientele. If the clientele cannot understand what it is that the institution is trying to tell them, then it's the responsibility of the institution to make it clearer.

If we say, as has been said, that the professionals should not interpret National Assessment Program data but the people should interpret it for themselves, I think that's a copout. It is the responsibility of the professional who collects the data to communicate effectively to the public as to just what it means and what it doesn't mean. Perhaps then we no longer would have school superintendents in Montgomery County, Maryland, or Oak Park, Illinois, or other suburban programs gleefully displaying their achievement test results to the newspapers, while at the same time school administrators in Washington, D. C., Chicago, Illinois, and Detroit, Michigan, are trying desperately to hide the results of similar test information.

There are reasons other than good or bad school systems for those results, and it is the responsibility of experts in the measurement field to interpret this kind of situation to the general public.

In my three years in the Office of Education I have rarely if ever had a communication with Congressmen, individually or as a committee, in which they seemed interested in knowledge for its own sake. There are few detached observers where power is dispensed. They were always interested in knowledge that would support or attack a point of view that they already had. I subscribe fully to Dr. Blake's point of view that we have to become much less naive about how the information we collect is being used in a public policy sense. We are in the middle of social turmoil, and we had best gain more insight and practice on how to comport ourselves under these changed circumstances.